

Benefits of increased public participation

US research, industries and citizens have something to gain from European experiments in public participation in regulation. So do those in Europe.

It would be absurd to suggest that science can be conducted democratically. It would be no less ridiculous to suggest that consideration of the impact of science and the regulation of its applications should be left entirely to the discretion of scientists and those who fund or employ them. But few governments of scientifically developed countries are sufficiently aware of the public's significant interest (in both senses of that word) in research to respond with the required commitment and resources.

It can be useful to distinguish, somewhat artificially, between active stakeholders associated with a particular issue and concerned publics. Stakeholders may include industrialists, investors in the stock market, food retailers, doctors, government ministries, farmers, lawyers, learned societies, publishers, the media, anti-biotech and green lobbyists, and disease sufferers' organizations. Publics have no immediate stake in the issue, but know that it will have an impact on the society in which they live and would willingly grasp an opportunity to have a voice.

Stronger foundations

A greater effort spent in giving non-stakeholder publics a voice and a demonstrable influence on the regulation of science and technology would leave scientific research, and industries built on it, in a healthier state. Not because society will necessarily be more accepting of a particular development as a result, nor because the public will necessarily be more scientifically educated; by virtue, rather, of the greater extent to which irrationality or vested interests can be countered and a foundation of greater public confidence and mutual trust established. Above all, the goal should be to build greater anticipation of progress and a stronger foundation for public consideration of, and beyond, the science. There is a prospect then that public discussion and regulation could be that much less likely to be held hostage to the interests of those who shout both loudly and misleadingly or those who wield disproportionate influence behind the scenes.

An inevitable objection to such a move is that the public lack the education and knowledge to assess science. In some complex topics this is a justified concern. But many scientists and industrialists who have been involved in public participatory events such as consensus conferences, or in the public consultative process recently established at the US National Institutes of Health, remark on the (to them) unexpected way in which the public representatives have little difficulty in absorbing key aspects of science in the discussion of ethical or human issues. They also welcome the ability of such lay panels, given the power to run the process and select their own witnesses, to penetrate screens of rhetoric and selective evidence.

Although some of its key decisions on regulation involving, for example, the Food and Drug Administration have been made too discreetly, the United States leads the world in providing information to the public on the web, and the extent to which it invites stakeholder consultation. A 1997 presidential commission on environmental-risk management made much of the way information to and consultation

of stakeholders had led to examples of good development while a lack of them had resulted in decisions that foundered.

But organizations in the United States can enhance their considerations of science's impact by a more systematic involvement of the public. Much of the development of techniques for public participation has occurred in Europe (for an overview of those techniques and examples around the world see www.oecd.org/dsti/sti/s_t/biotech/act/consultations.htm, and also the October 1999 special issue of *Science and Public Policy*). The pioneers of public participation have been the Danish Board of Technology, which has organized consensus conferences and scenario workshops for many years. The purpose of the board (see <http://www.tekno.dk/eng/index.htm>) is to help the Danish parliament be aware of the full range of issues associated with a particular technological development. The lay panel is briefed on those issues and, over a four-day period, can question anybody it wishes from research, industry and lobby groups about the issues, and writes a report (GM crops is one recent example) which is made freely available on the web. This panel represents the public interest in a way that stakeholders do not. The achievement of a consensus ensures at least that they have had to think through the issues in the required depth from a practical point of view.

The applicability of this model to other countries has been questioned on the grounds that, like the Netherlands, Denmark has a political culture that is unusually consensual. But that misses the point, which is that these conferences are not intended to reach policy decisions, and need not be applied at a national level. They are intended to shed light on citizens' concerns about key aspects of science and of issues surrounding it. To enlarge its consideration so as to include the wider public view, organizations in the United States would do well to build on such techniques.

New resources

Europe, too, should do more. In the United Kingdom there is a new set of bodies well placed, in principle, to incorporate such techniques into their formal responsibilities — the Human Genetics Commission, the Agriculture and Environment Biotechnology Commission and the Food Standards Agency. They have yet to prove themselves, and in their make-up are more detached from the interests of research and industry than were their predecessors. They should be given resources not only to provide high-quality and prompt contextual information for the media and the public, but also to develop public participatory studies of potentially contentious issues. Other countries in Europe — France and Germany — have a similar need, but show less readiness to move in this direction.

If they and the United States can draw more on the detached judgement of such panels, the process by which science can benefit society should improve. There is a need to adopt a more inclusive approach in grappling with the implications of increased scientific knowledge. Given what is sometimes at stake for economies and for citizens, there is no good excuse for not doing so. ■





Call of the wild
Bonobo research
threatened by civil
war in Congo
p262



Drug culture
Third World AIDS
treatment gets a
boost
p263



Free again
Environmentalists
score victory over
neem patent
p266



Helical hopes
Dresden unveils
grand plans for
biotech centre
p268

Europe agrees to boost Internet networks used by researchers

Declan Butler

The 15 member states of the European Union have approved an 80 million euro (US\$73 million) upgrade of Europe's research Internet networks. The move will provide the infrastructure needed to begin work on the concept of an advanced research computing 'grid'.

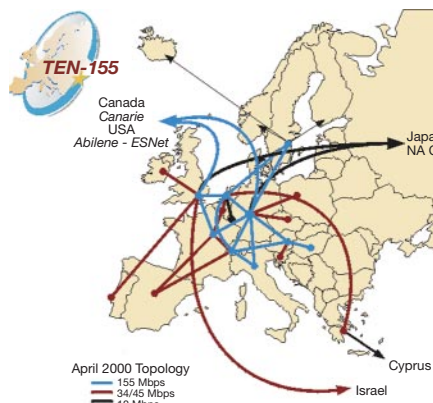
A bid for preparatory work on a research grid was submitted to the European Commission last week by the coordinator, the European Laboratory for Particle Physics (CERN), with five other national partners, including France's National Centre for Scientific Research (CNRS), Britain's Particle Physics and Astronomy Research Council (PPARC), and Italy's National Institute for Research in Nuclear and Subnuclear Physics (INFN), along with 15 other associate organizations, including IBM. EU investment is likely to be announced at a summit of political leaders in Lisbon next month.

The call for a proposal — unusual in that it asks for a single application from the European research community — represents a green light for the Géant project. This has been proposed by Dante, the Cambridge-based coordinator of 31 national research networks in Europe, as the successor to the current Ten-155 European research network.

Géant would boost the network's capacity from an average of 155 megabits per second to 2.5 gigabits per second this year, and to 10 gigabits per second and higher within four years. Member states fund the pan-European research network jointly, so the upgrade "will happen fast and be focused," says one commission official in Brussels.

The new commission funding stipulates that the network must operate at the same speed throughout Europe, even in poorer countries. "There is a need to avoid partitioning the European research infrastructure fabric and weakening European integration," says one commission document.

Dai Davies, general manager of Dante, is optimistic that this requirement can be met. "While national infrastructures are not of the same quality, this can be offset somewhat by balancing costs so that a part of the total are shared irrespective of geography," he says.



Ten-155: the current European research network.

Further support for research computing infrastructure may come from the Lisbon summit, which is likely to approve a broad commission proposal to boost the knowledge-based economy known as eEurope.

One commission official says that eEurope will extend Géant by organizing gigabit-per-second connections to the network from all campuses and research institutes in Europe. "The objective is to build a completely optical network throughout all Europe, including poorer countries, before 2005."

EU funding is likely to be of the order of 100 million euros. The overall cost could be over 1 billion euros, and would require national and regional agreement. The European Investment Bank — an autonomous body set up to further European integration by promoting EU economic policies — is said to be interested. The EU is likely to approve spending a further 50 million euros to build fast connections linking Géant to networks across the world.

Liberalization of EU telecommunications markets has prompted a 20-fold reduction in costs over the past two years, according to Davies. He adds that progress in extending into central and eastern Europe has been easier than in Portugal and Greece, where telecommunications operators are reluctant to abandon monopolistic tendencies.

The quality of infrastructure now in sight means Europe can start work on a more advanced notion of research computing, the so-called 'grid'. The idea is that scientists would plug into the network and immediately have at their disposal the entire processing resources of the network.

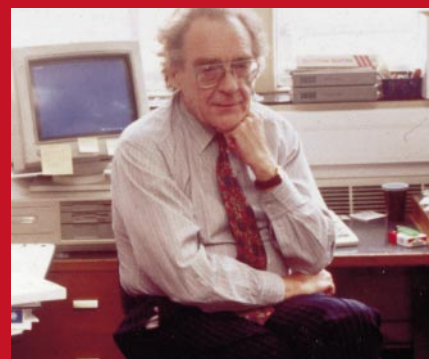
The term 'grid' comes from the analogy with the electricity grid, where one can plug in an appliance and the power needed is

A first Royal Society honour

Sir John Maddox (right), editor of *Nature* for almost 23 years before his retirement in 1996, was last week elected the first honorary fellow of Britain's Royal Society in London.

The position was introduced three years ago, and is intended to recognize individuals "who have rendered signal service to the cause of science, or whose election would significantly benefit the society by their great experience in other walks of life".

Maddox, a theoretical physicist by training, said on Monday that he was "delighted" at his election. "One can ask, is



this gamekeeper turned poacher, or poacher turned gamekeeper?" he quipped.

always there. The complex handling of data and jobs between centres would be hidden from the end user by a new generation of software known as 'middleware'.

CERN's \$1.8 billion Large Hadron Collider (LHC) is an ideal test bed for the grid, as it will generate 7 petabytes (10^{15} bytes) of data every year when it comes online in 2005. This information — requiring a thousand times more computing power than CERN can currently deliver — must be delivered to thousands of users in more than 40 countries. Using the grid concept, the Internet would function like a single computer and database rolled into one (see *Nature* 404, 213; 2000).

Political support for the programme is already high. Britain has pledged £100 million to grid computing. John Taylor, director general of the UK research councils, told research ministers in Lisbon last month that the grid represented a fundamental shift in computing, and would allow global teams of scientists to access large-scale computing resources and data collections distributed across many institutions and countries.

A \$500 million five-year grid effort is already under way in the United States, involving 50 research centres coordinated by the National Computational Science Alliance.

Preliminary discussions suggest that Europe's grid effort is likely to "match or exceed" that of the United States, says Robin Middleton, a scientist at the UK Rutherford Appleton Laboratory and the PPARC representative of the CERN proposal.

The CERN project is seen as a forerunner for similar proposals from other disciplines with similar computing needs, such as bioinformatics and astronomy. ■

Congo war increases threat to bonobo research

Asako Saegusa, Tokyo

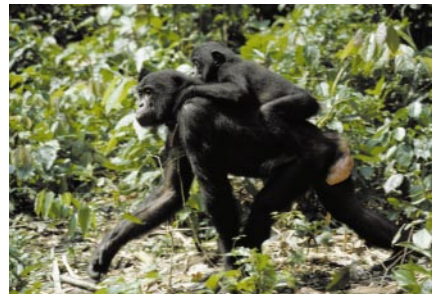
Future research on the rarest and least-known species of great ape is being seriously endangered by the intensifying civil war in the Democratic Republic of Congo, according to leading primate researchers.

They believe that threats to the wild population of bonobos (*Pan paniscus*), or pygmy chimpanzees, could jeopardize ongoing studies of their life histories and evolution. Work following up suggestions that bonobos might be useful in research on the origins of AIDS could also be harmed.

Wild bonobos are found only in Congo. Their population was estimated at around 10,000 in 1996, but is thought to have halved over the past few years as a result of habitat loss and an increase in hunting for their meat. The recent outbreak of armed conflict within core bonobo habitats is believed to have reduced their numbers still further.

"Ongoing [great ape] conservation programmes are subject to severe disruption all too often," says Richard Wrangham, a Harvard University primatologist. He notes that more than two-thirds of 23 protected areas containing great apes have been disturbed by military conflicts during the past ten years.

"Since bonobos were the last of the great apes to be studied, their life history is yet to be understood," says Takayoshi Kano, a primate researcher at Kyoto University, who launched a pioneering study on wild bono-



Bonobos: endangered on more than one front.

bos in 1973. "Even if bonobos do survive the ravages of war, their population will have been disrupted, and beginning a fresh study on a different group would be an unthinkably daunting task."

As human contact with bonobos increases with accelerated hunting — which often involves primitive butchery — there is growing concern over the risks that the apes could transmit disease to humans. Particularly in the light of the recent discovery that primate retroviruses can be transmitted to humans relatively easily through hunting.

Recent findings that HIV-1 might have arisen from chimpanzees in central Africa have triggered interest in bonobos as a research subject, although SIV — the ape equivalent of HIV — has not been detected from the handful of bonobo samples that have been tested so far. ■

NIH takes charge of chimps infected in experiments

Paul Smaglik, Washington

The US National Institutes of Health (NIH) has taken over responsibility for 288 chimpanzees infected with HIV and hepatitis C as a result of research projects. The chimps are currently housed at the Coulston Foundation in Alamogordo, New Mexico.

Coulston, a private research laboratory, has been criticized by animal-rights groups for the conditions under which its animals are kept, and has been chastised by the US Department of Agriculture (USDA). But it is still in the running to bid for contracts to care for the animals.

Coulston, which was founded by Frederick Coulston, lost two such contracts last year after USDA investigations ruled that the facility had violated the Animal Welfare Act. USDA inspectors reported that Coulston's chimp housing was dirty, infested and poorly ventilated. According to



Coulston: his lab can still bid for contracts.

the department, these conditions contributed to the deaths of three chimps.

The loss of the contracts — worth \$10 million over six years — put the Coulston lab in financial trouble and cast doubts over its future (see *Nature* 398, 644; 1999). Its financial situation "has become very critical", says John Strandberg, NIH director of comparative medicine.

But a successful bid by Coulston could help keep the organization solvent. Bids will begin in June and will undergo peer review. Don McKinney, a spokesman for Coulston, declines to comment on whether it will bid for the contract. However, he says he is "enthused" that NIH is supporting

care of the animals.

Most of the 288 chimps, which represent about half of Coulston's total chimp population, are housed at Alamogordo's Holloman Air Force Base, leased by Coulston. Of the 288, 25 remain in active clinical trials. The foundation will still own around 300 other chimps.

Strandberg says Coulston deserves to bid for the chimps' care, since conditions there have improved. "They have been intensively monitored," Strandberg says. "They've responded to many of the issues."

The California group In Defense of Animals, which opposes animal research, disagrees. It wants the permanent retirement of all chimpanzees used in government-sponsored research and opposes any involvement in their care by Coulston. "The NIH's deal does not accomplish either of these pressing goals," says Elliot Katz, the group's president. ■

Cheaper AIDS drugs due for Third World...

Phyllida Brown

The fight against AIDS in the developing world received a boost last week as five pharmaceutical companies joined forces with the United Nations to promote easier access to treatments. Public-health officials welcomed the move, but warned that cheaper drugs alone will not be enough to overcome the HIV epidemic in Africa.

In the same week, US president Bill Clinton issued an Executive Order conceding that African nations can use generic AIDS drugs rather than patented products from US-based companies.

The Joint UN Programme on HIV and AIDS (UNAIDS) and the five drugs companies — Boehringer Ingelheim, Bristol-Myers Squibb, Glaxo Wellcome, Merck and Hoffmann-La Roche — said on 11 May that they had begun a 'dialogue' on improving care and treatment for HIV and AIDS in developing countries.

Cheaper drugs will be one key component of the programme. For example, Glaxo Wellcome says it will now provide its anti-HIV combination Combivir (zidovudine and lamivudine), under specific treatment conditions, to developing countries at about \$2 per day, instead of the usual \$17 per day.

The UN agencies and other international bodies have been trying for at least a decade to get the price of AIDS drugs cut in Africa. To date, more than 80 per cent of all AIDS-related deaths have occurred in this continent. Recent calls to the industry from UN secretary-general Kofi Annan and UN agency heads for partnerships against AIDS appear to have been behind the companies' decision.

Peter Piot, head of UNAIDS, describes the move as "a promising step in a long-term process". But he warns of the need for "significant new funding" — for example, to ensure that African governments can provide more clinics and trained staff to administer treatment and to monitor patients for the emergence of drug-resistant strains of HIV.

The day before the announcement — and, industry spokespersons insist, entirely coincidentally — President Clinton issued his Executive Order. This prevents the US government from punishing any African government that makes or imports cheap generic AIDS drugs or medical technologies, such as HIV diagnostics, instead of patented products from US-based companies. The order lacks full legal status, but can only be overturned by another Executive Order.

Under the Agreement on Trade-Related Aspects of Intellectual Property Rights, governments can use compulsory licensing to allow the production or sale of products without the permission of the patent holder in exceptional conditions, such as public-health crises.



Feinstein: welcomes decision on generics.

In the past, when developing countries have tried to use this provision to make or sell cheaper generic AIDS drugs or diagnostics, the US Trade Representative has opposed them, using retaliatory trade tactics or other leverage.

The new Executive Order says that the US government must "refrain from seeking, through negotiation or otherwise, the revocation or revision" of legislation imposed by African governments to get cheaper AIDS drugs to their populations. But African governments must continue to protect US companies' intellectual-property rights.

Senator Dianne Feinstein (Democrat, California), who has been pushing for such a move, welcomed Clinton's decision. In a statement, she said that the order "will give sub-Saharan governments the flexibility to bring life-saving drugs and medical technologies to affected populations while ensuring that fundamental intellectual-property rights are protected".

But the American pharmaceutical manufacturers' organization, PhRMA, claims compulsory licensing weakens intellectual-property protection and so reduces the incentive for research into new drugs.

Alan Holmer, PhRMA's president, said last week that Clinton's move "sets an undesirable and inappropriate precedent". PhRMA says it recognizes that AIDS is a major problem, but argues that "weakening intellectual-property rights is not the solution".

...but South Africa voices reservations

Michael Cherry, Cape Town

South African government officials have reacted cautiously to a deal announced last week by five major companies to slash the price of HIV and AIDS drugs for developing nations (see above). They are concerned that this could undermine efforts to reduce the cost of other medicines.

Patricia Lambert, adviser to South African health minister Manto Tshabalala-Msimang, has said that any effort to bring down drug prices should be welcomed. But she added that concessions for obtaining cheaper AIDS drugs should be applied to other medicines as well.

Lambert was also reported by Dow Jones newswires as saying that "if this offer is attached to a condition that governments like South Africa should not pursue generic substitution, parallel importing and compulsory licensing, then it is not genuine and unacceptable".

But spokespersons for the pharmaceutical industry have denied any formal link between intellectual-property issues and the companies' deal. The deal, announced in Geneva last week, involves Boehringer Ingelheim, Bristol-Myers Squibb, Merck, Hoffmann-La Roche and Glaxo Wellcome.



Further AIDS? Concessions 'should be applied to other drugs too'.

Both Mirryena Deeb, chief executive of the Pharmaceutical Manufacturers' Association of South Africa, and Peter Moore, medical director in Glaxo Wellcome's Johannesburg office, said they are unaware of any formal link between the offer to reduce the costs of the drugs and patent protection.

Moore suggests that the deal should make the generic substitution of AIDS drugs unnecessary, and says there is no substance to Lambert's suspicions of a hidden agenda to protect the price of other drugs. "To the contrary, if this programme works with AIDS drugs, there are no reasons why it should not be extended to other diseases as well," he says. "But the position

will change in South Africa only when the concerned parties are prepared to sit around the table and trust each other."

Last week's announcement came against a backdrop of parliamentary hearings on HIV and AIDS in which health director-general Ayanda Ntsaluba revealed that he had been talking to the trade and industry department with a view to introducing compulsory licensing which would allow access to cheaper drugs.

Deeb called Lambert's statement "disappointing, as she is questioning the bona fides not only of the industry, but of the World Bank and UNAIDS", the agencies that brokered the deal. "If it is true, her statement is an ungracious response to a genuine attempt to make a difference."

Geneticist from Baylor named as new head of UK's Sanger Centre

David Dickson, London

The Wellcome Trust announced this week that a mouse geneticist is to be the next director of the Sanger Centre. Based outside Cambridge, England, and founded jointly by the trust and the Medical Research Council, the centre is one of the world's leading sequencing facilities.

Allan Bradley has spent the past 13 years at the Baylor College of Medicine, Houston, and is currently Cullen Professor of Human and Molecular Genetics. He is also a Howard Hughes Medical Institute investigator. He will take over the leadership of the Sanger Centre from sequencing pioneer John Sulston, who has headed the centre since it was set up in 1992.

Bradley studied under geneticist Martin Evans at the University of Cambridge before taking up a post at Baylor in 1987. His main research interest is in elucidating the developmental genetics of the mouse, in particular by isolating stem-cell lines from pre-implantation embryos.

"Remarkable opportunities are arising from genome sequencing programmes to progress basic biology and its applications in medicine," says Michael Dexter, the head of the Wellcome Trust. "We are delighted that Allan Bradley is joining us and helping to drive forward the trust's mission."

The main task facing Bradley will be to map out the 'post genome' phase of Sanger's development, now that the sequencing of the human genome is nearing completion. His background suggests that both mouse genomics and stem-cell research will — as with the US National Institutes of Health — be high on the agenda. ■

Researchers take a gamble on the human genome

Paul Smaglik, Washington

Biologists are taking bets — literally — on the number of genes in the human genome. Enterprising attendees at the annual Cold Spring Harbor genome meeting last week opened a book, taking bets at \$1 a time. The results will be known in three years time, when sequencing is completed.

The spread of bets placed so far — from 27,462 genes at the low end to 153,478 at the high — represents two very different approaches to gene counting. Techniques that extrapolate from manually annotated portions of the parts of the genome that have already been fully sequenced are yielding estimates of around 35,000–40,000 genes. Those that use computer algorithms to scour random expressed sequence tags (ESTs) from the whole genome predict 100,000 or more.

Bets cost \$1 this year, \$5 at next year's meeting and \$20 in 2002. At stake — beyond a cash prize that will be awarded at the 2003 meeting — is an improved understanding of the complexity of the genome, and the relative importance of genes and regulatory regions. There is still confusion over how to count genes that don't code for protein, and uncertainty about what biological roles these pieces of DNA play.

Perhaps reflecting the absence of widely expected announcements on new sequencing milestones, the issue became the focus of heated discussion at Cold Spring Harbor last week. "Sequencing is like digging a gold mine," says John Quackenbush, a computational biologist with The Institute for Genomic Research (TIGR) in Rockville, Maryland. "How much gold is there to find?"

Quackenbush's own prediction of 120,000 genes, to be published in the June issue of *Nature Genetics*, is at the high end of the scale. DoubleTwist (formerly Pangea), a bioinformatics company in Oakland, California, last week announced that its own algorithms predicted a total of 100,000 genes.

Both use gene-hunting programs and DNA and protein homology searches to find candidate genes in GenBank. But each uses different software. Another genomics company, Incyte, predicts over 100,000 genes, but has used its proprietary EST database, rather than the public GenBank (see *Nature* 401, 311; 1999).

Tim Hubbard, who heads Ensembl, an automated annotation program similar to TIGR's and DoubleTwist's, has doubts. "Automatic annotation over-predicts the number of genes," says Hubbard, who is based at Britain's Sanger Centre. "This is due to false positives and cases where multiple genes are annotated when there is really only one."

The Ensembl total mirrors that of groups led by Philip Green, of the University of Washington, Seattle, and Jean Weissenbach, of the Centre National de Séquençage at Evry, France, both of which have separate papers appearing in June's *Nature Genetics*. Each used separate techniques that closely scrutinized a portion of the genome before scaling up their counts.

Green's group used either the curated, annotated genes from chromosome 22 or a filtered set of near full-length mRNA sequences from Genbank as a starting point. Weissenbach's team compared the bacterial artificial chromosome clone ends of a pufferfish genome, with a known, curated set of protein-encoding genes, then used algorithms to compare protein coding regions from the fish's genome with a human sequence.

Quackenbush notes that extrapolation has its weaknesses, too. For example, chromosomes 22 and 21 — whose sequence is published this week (see page 311) — are similar in size; but chromosome 21 has 225 genes, compared with 545 on chromosome 22. Using either total on its own to estimate the number of genes in the genome could be misleading.

Francis Collins, director of the US National Human Genome Research Institute, notes that totalling the genes in chromosomes 21 and 22, then scaling that figure up to account for the size of the whole genome, results in an estimate of 40,000 genes. That's close to the figure arrived at by the extrapolators and Ensembl. But he cautions not to put too much faith in this approach. ■

♦ <http://www.ensembl.org/genesweep.html>

Los Alamos labs are safe from fire

Rex Dalton, San Diego

The Los Alamos National Laboratory in New Mexico was reported this week as being out of fire danger, following a blaze that burned 43 square miles of nearby forest.

The homes of 216 laboratory employees — including many scientists — in nearby communities were destroyed in the fire. Officials say the blaze was started as a controlled burn by forestry agents, but that it then got out of control.

Laboratory employees began returning to work on Monday, after the facility had been closed and largely evacuated for nearly a week, leading to about \$3.5 million a day in operational losses. Although there was no



damage to any facilities housing nuclear material, two old, isolated structures from the Manhattan Project were destroyed. ■

Gene bank to offer family album of mammals

Natasha Loder

The world of ecology and evolutionary biology is about to benefit from high-throughput DNA sequencing. A joint project between the Zoological Society of San Diego and Amersham Pharmacia Biotech (AP Biotech) is set to sequence DNA from every family of mammals.

The two organizations say the result will be the world's most comprehensive database of mammalian mitochondrial DNA. They hope the database will become a major tool for research on the evolution and conservation of mammalian biodiversity. It should also be useful to researchers working on the genetic bases of species' behaviour, biology and evolution.

The collaborators say they will publish regular progress reports, and that the results of the sequencing will be made public "quickly".

Over the past 25 years, the society — which runs San Diego Zoo — has maintained a collection of cell lines known as the Frozen Zoo. This means that it already has DNA samples from over 100 of the 150 or so mammalian families.

The task of preparing the samples for sequencing is in the hands of AP Biotech company Molecular Dynamics, based in Sunnyvale, California. Robert Feldman, the production, sequencing and collaborations manager, says that his company is keen to develop the experience of broad exposure to a wide and diverse set of DNA sequencing projects. He declined to comment on the cost of the scheme.

Mitochondria are the structures in a cell



responsible for most of its respiration and energy production. They are popular with researchers interested in evolution at the molecular level, because they have a single, small (around 16,500 base pairs) chromosome which is passed down from mother to offspring. This makes them easier to analyse than full genomes.

Mitochondrial DNA contains a number of genes that evolve at different rates. This means mitochondria can be used for evolutionary analyses over different time scales.

Oliver Ryder is an adjunct professor of biology at the Center for Reproduction of Endangered Species at the Zoological Society of San Diego, a collaborator on the project and a leading figure in conservation genetics. He says that variation in mitochondrial DNA — even at the family level — will be useful for interpreting the results of conservation

genetics studies, and for analysing evolutionary differences between species.

"There are many examples where mitochondrial DNA has provided profound insights into this area of whether populations [of animals] are the same or significantly diverged," he says. For example, variation in parts of the mitochondrial DNA of cheetahs and gorillas has helped researchers understand the evolutionary diversity of regional populations and subspecies.

"Differences between families will help us understand the difference between genera and species. A lot of work is done on the tips of the evolutionary tree in the absence of the trunk and the limbs," he adds.

Ryder suggests that the database may also bring a deeper understanding of how mitochondria evolve. At a broader level, the database will offer a "scaffold of data" for interpreting mitochondrial DNA diversity.

Ryder and colleagues recently argued in favour of starting DNA banks to store the world's genomic biodiversity. They believe that collections started now would be useful in the future for reasons we cannot anticipate (see *Science* **288**, 275–277; 2000). He points to the new project's use of the society's Frozen Zoo, as an example of this.

Ryder thinks the database is timely, "given the capacity in genomic sciences and the opportunity in taking the focus away from medical and therapeutic uses for a single species — humans". Feldman agrees: "This project takes the genetics of endangered species into the genomic era. That for me is very exciting," he says.

► <http://www.sandiegozoo.org/cres/>

Japan calls for open access to human genome data

Robert Triendl, Tokyo

The genome science committee of Japan's Council for Science and Technology is set to call for the open release of human genome data. The council coordinates Japanese research policy and is chaired by the country's prime minister.

The council's forthcoming declaration says that, to assure scientific progress, "all data and results" from human genome research should be made public. This is expected to include complementary DNA and single-nucleotide polymorphisms.

Last month, the director-general of the Science and Technology Agency, Hirofumi Nakasone, called for the creation of international rules on the release of human genome data. But although the declaration will probably cover all publicly funded

research, observers in Tokyo point out that Japanese funding agencies do not usually lay down policies for data release.

Earlier this year, the Research Association for Biotechnology, which administers biotechnology research for the Ministry of International Trade and Industry (MITI), announced the release of sequence data for 2,200 human genes — but only after the Japanese Patent Agency said that it would not approve patents for sequences that are not complete and whose functions are unknown.

The data have been produced by a MITI-sponsored consortium of private companies, which aims to sequence human full-length complementary DNA.

Patents for the genes are pending. The release of the genes' sequences will prevent overseas companies from filing patents on

them, and the period granted for revising the applications will give the companies involved some time to identify gene functions.

A spokesperson for the Research Association for Biotechnology said that although the cDNA consortium is funded entirely by public money, there are no clear rules for the release of data. Results have to be reported to the New Energy Development Organization, a MITI funding body, within 60 days of a project's completion.

Opposition from MITI has made it difficult for Japan to formulate a position on the release of human genome data. Akiyoshi Wada, director of the RIKEN Genomic Science Center, says that he recommends the regular and timely release of data to his staff, but the lack of general rules has made it difficult to draft policies for his institute.

Greens persuade Europe to revoke patent on neem tree...

Ulrike Hellerer, Munich

K. S. Jarayaman, New Delhi

A global coalition of environmental groups is celebrating a major victory in their fight against what they call 'biopiracy'. Last week, the European Patent Office (EPO) revoked a controversial patent on the use of antifungal agents extracted from the neem tree.

The patent was granted in 1994 to the US Department of Agriculture and the multinational agribusiness corporation W. R. Grace. But it has been fiercely contested by environmental groups, to whom it is a symbol of western multinationals' unacceptable exploitation of indigenous communities.

The neem tree (*Azadirachta indica*), which translates as the 'free tree', is a member of the mahogany family. It is indigenous to the Indian subcontinent, where it has been used in agriculture, medicine and cosmetics for centuries.

The EPO has received 51 patent applications for 'inventions' based on the neem tree, and has so far granted 11. Around 90 patents

exploiting the tree have been granted worldwide.

The coalition fighting the neem patent — whose motto is 'Free the Free Tree' — is coordinated by the Green party in the European parliament. It includes the Indian Research Foundation for Science, Technology and Ecology, an influential environmental group, and the International Federation of Organic Agriculture Movements (IFOAM), a pressure group for organic farming.

In applying for the patent, Grace had argued that it had used an extract of the tree's seed to make a new fungicide. But the coalition claimed that Grace's patent was not sufficiently novel, as Indian farmers have used this fungicide for decades.

The EPO's opposition board accepted this interpretation, because Ajay Bio-Tech — an agrochemical company in Pune, India, that has been making a neem-based fungicide since the early 1980s — produced documents to prove that it had conducted the relevant field trials.

Protests against the patenting of neem



Free Tree freed: protesters celebrate their victory outside the Munich patent office.

products began in India in 1993, with one demonstration in Bangladesh numbering half a million people. The Free the Free Tree coalition filed its opposition to the patent in 1995. It also launched a petition in India in 1994, and presented the EPO with 500,000 signatures during last week's hearing.

Linda Bullard, president of IFOAM, says that the coalition lacks the resources to dispute the other ten neem-related patents granted by the EPO. But she hopes that the precedent set by last week's ruling will

CHRISTINE STRUB

influence decisions on pending patent applications — a possibility the EPO concedes.

Ragunath Mashelkar, director-general of the Council of Scientific and Industrial Research in New Delhi, the coordinating body for India's government research laboratories, says that the decision may not have a direct influence on all cases. But he welcomes the withdrawal of the patent "in the sense that it sensitizes the world against biopiracy".

Beth Burrows, the president of the Edmonds Institute, a non-profit environmental organization near Seattle which also supported the objection, says that "broadcasting the message that the patenting system facilitates biopiracy, the appropriation of biological resources and knowledge from southern countries, was one of the coalition's main aims". She adds that the neem patent "robs the poor of the last resources that can ensure their survival".

W. R. Grace is reserving comment. It has two months to challenge the EPO's decision. Abhay Phadke, managing director of Ajay Bio-Tech, believes the decision is a step towards preventing W. R. Grace monopolizing the market in neem seed. But he points out that W. R. Grace remains in a powerful position, because it has similar patents filed in the United States.

Furthermore, the EPO's decision may not be generally applicable. Birgit Reiter, an expert in pharmaceutical and patent law at the German Association of Research-Based Pharmaceutical Companies, says that last week's technical decision will have no direct influence on the pharmaceutical industry's interest in exploiting natural products, as it was not based on a general principle about the rights to such harvesting.

But she adds that if it becomes more diffi-

cult to patent natural products, "pharmaceutical companies will have to think again about whether the expense of further development and clinical trials will pay off".

The environmentalists are now happy to relax for a while. "The whole issue, and what to do next, will be discussed again at the meeting of the Biodiversity Convention in Nairobi this week," says Bullard. "The decision will give new impetus to the discussions." ■

...as India pushes ahead with plant database

Indian science minister Murlu Manohar Joshi greeted last week's revocation of a European patent on the neem tree — which the Indians have traditionally used in medicine and agriculture — almost as if it were a military victory. But his advisers know that piecemeal success will not solve the problem of 'biopiracy'.

With this in mind, India has announced plans to create a digital database of its traditional knowledge. This will be included

in the patent classification system of the Geneva-based World Intellectual Property Organization (see *Nature* 401, 413; 1999).

The database will be available to patent offices worldwide — especially in the United States and Europe — so that data on any Indian plant can be obtained before patents are issued. "This is the only permanent solution to prevent patents from being issued for non-original inventions in our traditional system," says

Ragunath Mashelkar, director-general of the Council of Scientific and Industrial Research in New Delhi.

Mashelkar told *Nature* that the documentation of 90 indigenous plants with medicinal or industrial uses has already begun — a complete electronic database on these will be available in between six months and one year. All major plants are expected to be covered in the next two years.

K. S. J.

Science academies pool their wisdom on global issues

Tokyo An international group of science academies is today setting up an organization to bring scientific resources to bear on issues of international concern, such as food security and the control of emerging diseases.

They agreed to form the InterAcademy Council (IAC) at a meeting in Tokyo last weekend. The idea was first mooted by Bruce Alberts, president of the US National Academy of Sciences. The council will have its headquarters at the Royal Netherlands Academy of Arts and Sciences in Amsterdam, and will function as a formal advisory body, funded on a project-by-project basis by international agencies.

"The burgeoning threats to global sustainability call on the world's scientific community to do more," says P. N. Tandon, a former president of the Indian National Science Academy and co-chair of the InterAcademy Panel on International Issues that established the IAC.

In a statement welcoming the creation of the IAC, Kofi Annan, the UN secretary-general, said that the complex issues confronting the international community "demand a bold vision for action involving a wide array of international, national and local institutions".

Germany's mobile gene school goes up in smoke

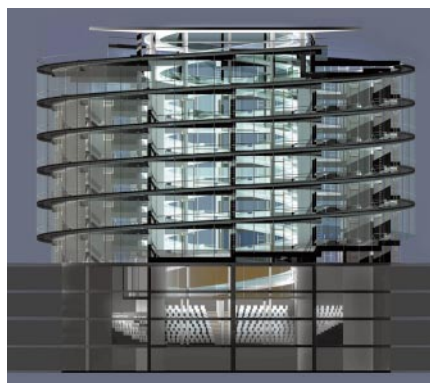
Munich Germany's new GeneTech-mobile, launched last month as a central focus of the federal research ministry's information campaign on gene technologies, is in ashes.

The mobile laboratory and information centre was completely burnt out last week after a visit to a school in the state of Hessen. Police suspect arson, but no group has so far claimed responsibility.

The presumed attack followed a heated public discussion about genetic engineering in the school in Giessen. Last year a similar vehicle, the BioTech-mobile, was burnt out in Giessen, which lies close to an area where field trials on GM crops have been repeatedly destroyed. The GeneTech-mobile will be rebuilt at a cost of DM1.1 million.

Dresden has designs on biotech research

Munich The design of Dresden's new bioinformatics centre will be based on the DNA double helix (see picture). The building will be constructed in city-owned castle grounds near the river Elbe as part of the planned 'BioParc Dresden'. The BioParc will be jointly financed by the city of Dresden, the state of Saxony and the Klaus Tschira Foundation —



In a whirl: Dresden's planned biotech centre.

recently set up to support science by a founder of SAP, Germany's largest software company.

With the new centre, Dresden hopes to move into the premier league of biotechnology and genome research in Germany. Kai Simons, founding director of the new Dresden-based Max Planck Institute for Molecular Cell Biology and Genetics, is particularly happy about the initiative. A bioinformatics facility would harmonize perfectly with research carried out at his institute, he told a press conference this week.

Spain plans to increase access to information

Barcelona Spain is to spend Ptas 550 million (US\$ 3 million) over the next three years on the dissemination of science and technology. The programme, presented last week in Madrid as part of the National Plan on Research, Development and Technological Innovation 2000–2003, is considered to be a priority area and is targeted at researchers in both public and private non-profit institutions.

Activities to be supported include the creation and development of web sites in research centres — to act as access 'portals' to wide spectra of information — and surveys of public knowledge about and attitudes towards science and technology. The programme also includes support for science journalists, including the possibility of stays in research centres.

2003 US Mars mission down to two options...

Washington The US space agency NASA might send an orbiter or a lander to Mars in 2003 — but it can't afford both. By July, NASA expects to decide between either making up for the lost Mars Climate Orbiter, with a probe that will focus on the martian atmosphere and a search for water, or placing a mobile lander on the surface that would be able to travel up to 100 metres a day.

The lander — based on Cornell University's 'Athena' design — would land

inside Pathfinder-style airbags, but would be much more capable than the Pathfinder rover. Until last year's failures prompted a re-examination of the entire Mars programme, NASA was planning to send both an orbiter and a lander at each launch opportunity.

...as France stays on track for mission in 2005

Paris French research minister Roger-Gérard Schwartzberg voiced his support early this week for continued collaboration with the US space agency NASA over a mission to Mars, despite the recent setbacks to US Mars projects. Scheduled for 2005, the joint mission would aim to bring back rock samples.

In an interview published in *Le Figaro*, Schwartzberg said that France would pursue its investment in the Mars mission, estimated at FF2 billion (US\$279 million). "It's admittedly a costly programme, but we can't content ourselves to be a spectator when it comes to Mars exploration," he said.

Readers get access to Celera's data

Washington In landing its first academic customer, Celera Genomics may have found a way to balance its commercial interests with the public's concerns about access to data. For the next five years, readers of papers by researchers from Vanderbilt University, Tennessee, that are based on Celera's genome will be able to look at the relevant section of the commercial database for free.

The university can also seek available patents without paying Celera 'reach-through' rights. But Celera remains opposed to database redistribution. "We are not free to copy the database in any way, shape or form," says Mark Magnuson, Vanderbilt's assistant vice-chancellor for research.

The cloned mouse that roared is silenced

London Cumulina, the first mouse to be cloned from an adult cell, has died — at the ripe old age of two years and seven months. The average mouse lives for only about two years, and Cumulina's longevity will fuel the debate over the effects of cloning on the ageing process. A university statement said that, apart from a skin tumour — common in ageing mice — she remained healthy and active until shortly before her death, "while sleeping".

Created by a team led by Teruhiko Wakayama and Ryuzo Yanagimachi (*Nature* 394, 369–374; 1998) of the School of Medicine at the University of Hawaii, Cumulina was born on 3 October 1997. The impact of the research led to her being dubbed 'The Mouse that Roared'.

Reservoirs dog AIDS therapy

Pools of latent HIV, lurking in the cells of infected people, remain untouched even by powerful drug combinations. Paul Smaglik reports on how this is forcing researchers to rethink their strategies for fighting the virus.

"Depressing." That's how Robert Siliciano, a virologist at Johns Hopkins University in Baltimore, Maryland, describes the implications of his work on 'reservoirs' of HIV. Believed to consist of viral DNA held safely inside 'resting' host cells, the stubborn resilience of these reservoirs has made many researchers pessimistic about the chances of ever completely eliminating HIV from the bodies of infected people.

As it has become clear that HIV reservoirs persist in the face of the most potent drug cocktails available, many AIDS researchers are shifting their plans of attack. "The concept of a reservoir is of paramount importance to the whole philosophy of what we wish to accomplish therapeutically," says Tony Fauci, director of the US National Institute of Allergy and Infectious Diseases in Bethesda, Maryland. The problem is that reservoirs provide a means for the virus to bounce back from assault by drugs, potentially re-emerging in a drug-resistant form.

A few researchers still hope that it will prove possible to drain the reservoirs by finding new, more powerful drugs. But potent drugs tend to have toxic side effects and, given HIV's apparent ability to evolve resistance to any drug thrown at it, a fresh spotlight is being turned onto boosting the immune system's ability to fight back against the virus. If this can be done, it might be possible to control HIV infection, even if the virus can't be eliminated, allowing infected people to live long and healthy lives.

Combined forces

The current uncertainty is a far cry from the excitement of the mid-1990s, when some researchers dared to think that the right combinations of drugs might provide a cure for AIDS. In 1995, two groups — one led by David Ho, director of the Aaron Diamond AIDS Research Center in New York, the other by George Shaw of the University of Alabama at Birmingham — presented data suggesting that HIV engages in a massive, but relatively unsophisticated assault on the body^{1,2}. The virus targets CD4⁺ cells, a particular class of white blood cell involved in the immune response. From near the start of infection, the two teams argued, around a billion new viruses are produced each day. This rampant replication kills a similar number of the CD4⁺ cells. Over time, the

researchers suggested, the immune system loses its ability to replenish CD4⁺ cells at such a phenomenal rate. As a result, the patient's CD4⁺ cell count falls, leading to immune deficiency, opportunistic infections — and eventually death.

If this picture was correct, the implication was clear: hit the virus hard enough with drugs that stop its replication, wait for chronically infected CD4⁺ cells to die, and it might just be possible to eliminate HIV from the body. The introduction of drug cocktails known as highly active antiretroviral therapy (HAART) buoyed the mood further. HAART typically combines several drugs to block the action of two enzymes — reverse transcriptase and protease — which are central to HIV's cycle of replication and infiltration of new host cells. By 1996, clinicians monitoring some patients put onto HAART were reporting anecdotally that they could find no evidence of HIV in the patients' blood.

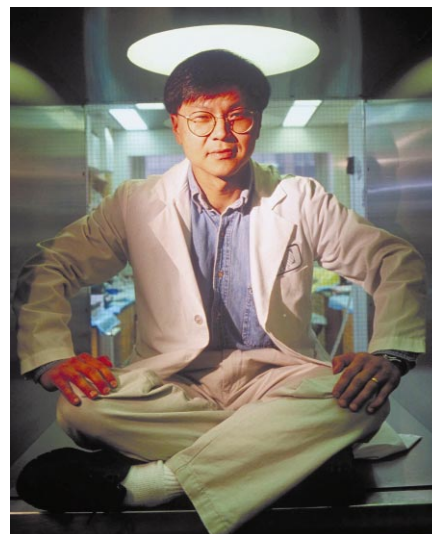
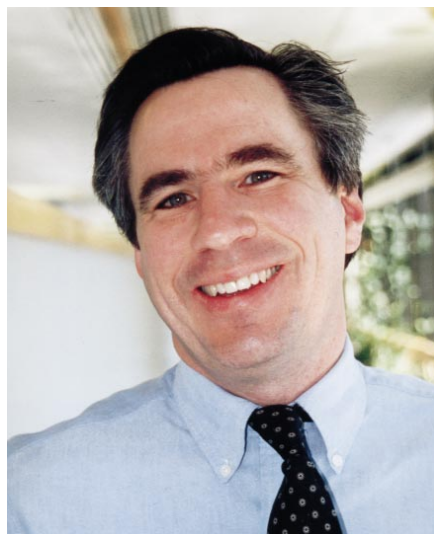
However, Siliciano's research was already throwing a spanner into the works. In 1995, before HAART came into widespread use, his team showed that a small number of infected CD4⁺ cells enter a 'resting' state in which HIV DNA integrates into the cell's chromosomes and simply sits there³. In these cells, the virus does not engage in the orgy of replication and destruction described by Shaw and Ho. In 1997, working in collaboration with Ho's team, Siliciano produced one of a flurry of papers showing that these reservoirs were still

present in patients on HAART⁴⁻⁶. Previous researchers hadn't detected the reservoirs because they didn't flush the virus out of resting cells. And late last year, researchers in Fauci's lab confirmed that, when HAART stops, HIV can bounce back from the reservoirs with a vengeance^{7,8}.

Phantom menace

The precise nature of the reservoirs remains a matter of debate. As well as resting CD4⁺ cells, there may be other cellular reservoirs that are not fully understood. But knowledge of how the reservoirs persist is advancing. Last year, for instance, Siliciano presented evidence suggesting that latently infected CD4⁺ cells divide slowly, copying their cargo of HIV DNA as they do so⁹. So although CD4⁺ cells eventually die of old age, their viral legacy lives on. In addition, other CD4⁺ cells might become latently infected because of a low level of ongoing viral replication.

Ho still hasn't given up on the idea of eradicating HIV using drugs alone. He compares a viral reservoir to a sink with a leaky plug and a dripping tap. In a best-case scenario, he suggests, patients who adhere strictly to their HAART regime might empty their reservoirs in about six years. Ho bases this conclusion on studies of eight individuals who began HAART soon after they were infected with HIV¹⁰, and other patients who were responding particularly well to



The eliminator: Ho (right) believes that HIV can be eradicated from the body, but Siliciano is sceptical.

HAART¹¹. In at least some patients, Ho concludes, it might be possible to turn off the tap completely. "It's still going to be difficult," he says. "I'm not trying to underestimate the task."

Mission impossible?

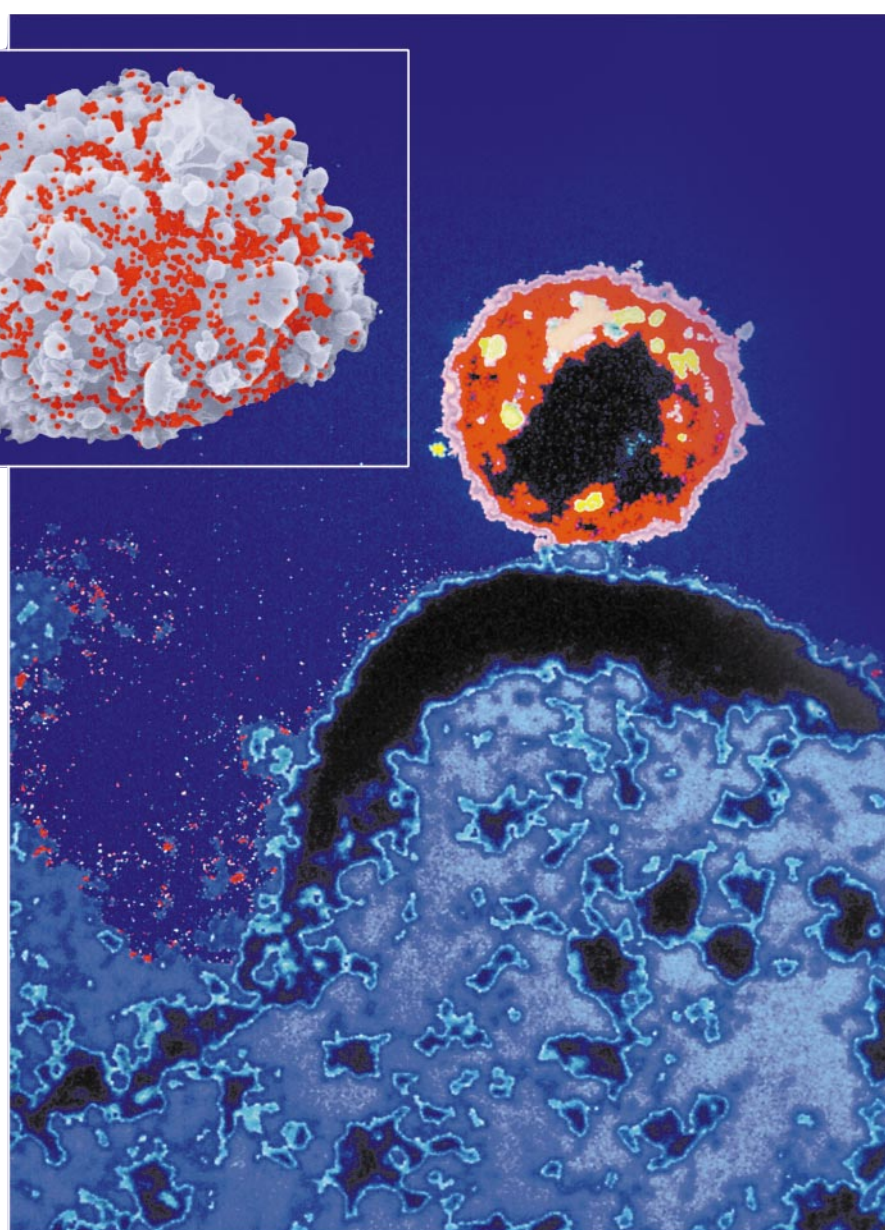
But many AIDS researchers don't share Ho's optimism about eradicating the virus. Frank Miedema, an immunologist at the University of Amsterdam, is particularly outspoken. "It's completely impossible to eradicate the virus with the current drugs," he asserts. Siliciano, who has tracked about 60 patients with good records of drug compliance over long periods, has found that the size of the reservoirs decays so slowly that they are likely to persist for life⁹ (see figure, overleaf). "It's possible that there is more rapid decay in a subset of people," Siliciano says. "We have not seen it."

Ho points out that many drug strategies remain untried. "We should not completely give up on the eradication idea until we have a good go at attacking with more potent combinations at all the sites available," he says. Drugs that inhibit protease and reverse transcriptase target HIV inside infected cells, Ho notes, while other drugs still in development aim to prevent HIV entering host cells. For instance, Trimeris, a biotechnology company in Durham, North Carolina, is developing two peptides that bind to gp41, a protein carried on the surface of HIV that is crucial for its fusion with the host cell's membrane. Meanwhile, other researchers are focusing on receptors found on the surface of cells susceptible to HIV, to which the virus attaches itself. Two of these receptors, called CCR5 and CXCR5, normally bind to chemokines, proteins that lure immune-system cells to diseased or damaged tissues. Drugs companies are trying to find chemicals that block these receptors.

The HAART of the matter

Many researchers doubt that any new drug strategy will prove much more effective than HAART at eliminating viral reservoirs. But even if drugs alone aren't the answer, they could be used in conjunction with strategies to boost the immune system. In theory, HAART could get the virus down to manageable levels, and therapies that boost the immune response might then be able to contain it.

Curiously, one approach involves stopping HAART for short periods of time. HAART is a mixed blessing, explains Bruce Walker, an immunologist at the Massachusetts General Hospital in Boston. "When you're put on HAART, your immune response drops down to lower levels," he says. Although viral replication surges as soon as patients are



Veiled threat: HIV (red) buds from the cells in which it is replicating (enlarged in main picture). When it does so it is vulnerable to drugs, but when it lies dormant it is untouched by therapy.

taken off HAART, it is bouncing back from a very low level, and the concurrent boost in the patients' immune response might just allow the virus to be recognized and controlled.

Walker has tried interrupting HAART in nine volunteers. "We are very encouraged by what we are seeing," he says. The results have not yet been published, but the patients showed variable increases in the numbers of circulating cytotoxic T lymphocytes (CTLs), the 'killer' cells that can destroy HIV-infected cells. And in some patients, there were signs that these cells were having the desired effect. In one individual, the levels of circulating virus remained low for several months after the second interruption of HAART.

One appealing aspect of the interruption strategy is that occasionally stopping HAART reduces the treatment's toxicity. But at the same time, interrupting therapy can

encourage the emergence of drug resistance. Stop HAART for too long, and the danger is that it won't be possible to control the virus once the treatment resumes.

Walker believes that, by itself, the interruption strategy is only likely to work in recently infected patients who were put on HAART as soon as their HIV status was known. Such patients are the 'best case', as the virus will have had little time to damage the immune system before the therapy begins.

Several teams are attempting the same strategy in chronically infected patients, who didn't begin HAART straight after infection. Published findings¹² and results presented at meetings are less encouraging than those obtained by Walker. "If there is an effect, it is only partial and in a subset of patients," says Giuseppe Pantaleo, an immunologist at the University of Lausanne in Switzerland.

Other immune-boosting treatments include infusing patients with interleukin-2 (IL-2), one of a family of proteins known as cytokines, which help coordinate our immune responses. Some AIDS researchers had hoped that intermittent treatment with IL-2, perhaps in association with other cytokines, might flush out the viral reservoirs by activating the resting CD4⁺ cells—rendering the hidden HIV susceptible to HAART. While the latest research from Fauci's lab suggests that it won't completely eliminate the reservoirs¹³, IL-2 does boost the CD4⁺ cell response. This could make it an important component of strategies to help the immune system regain control over HIV infection.

A shot in the arm

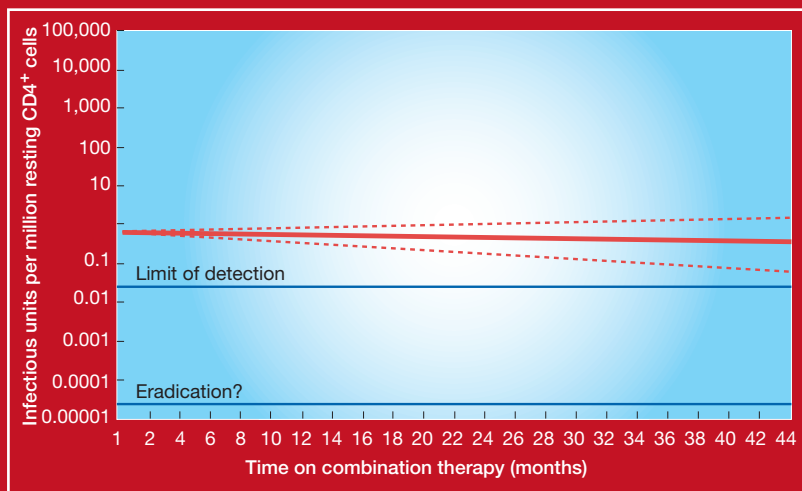
For now, however, the main growth area is in 'therapeutic' vaccine research. Rather than using vaccines in an attempt to prevent initial infection, the idea is to vaccinate HIV-positive people being treated with HAART, then halt the drug therapy. Hopefully, the vaccine will stimulate their immune systems sufficiently to bring the infection under long-term control.

A range of vaccines has been developed with the goal of preventing HIV infection—although none have yet achieved notable success. But even if vaccines can't stop people becoming infected, some might prove useful in a therapeutic context. Researchers at the Aaron Diamond AIDS Research Center have obtained intriguing results using a vaccine based on a canarypox virus, engineered to produce four HIV proteins. This is given in combination with recombinant gp160, an HIV protein involved in the virus's entry into host cells. "We have the first glimmer that there may be a role for therapeutic vaccines," says Marty Markowitz, the centre's clinical director.

Of 12 patients who received the canarypox/gp160 vaccine along with HAART, four were then taken off the drugs to see if their immune systems could control the virus. In two of these patients, levels of circulating HIV surged. But in the other two, the virus rebounded much more slowly. These two patients had a strong CTL response, while the other two did not. Markowitz acknowledges that the experiment, which has not yet been published, was small and had mixed results, but he regards it as a springboard for more research. Perhaps tinkering with the vaccine to produce more specific immune responses will do the trick, he suggests.

Another therapeutic vaccine, known as Remune, is undergoing trials in the United States, Europe and Thailand. Marketed by two companies in California, the Immune Response Corporation in Carlsbad and Agouron Pharmaceuticals of San Diego, this vaccine is based on whole HIV particles, stripped of a protein called gp120, and killed by irradiation and chemical treatment.

A life-long problem?



This plot, published last year⁹, shows just how resilient latent reservoirs of HIV can be. Researchers led by Robert Siliciano of Johns Hopkins University in Baltimore monitored the size of reservoirs in 34 patients receiving potent drug combinations over extended

periods. From the line of best fit for these data (shown above with 95 per cent confidence intervals), Siliciano and his colleagues calculated that, in a typical patient, it would take 60.8 years to eliminate reservoirs of HIV using current drug regimes.

Encouragingly, the vaccine causes CD4⁺ cells to proliferate. But so far, there is no evidence that the immune response it stimulates is sufficient to control HIV infection¹⁴.

Tit-for-tat tactics

"I think we will have to figure out a way to make these vaccines more immunogenic," says Fauci. For instance, it might be necessary to stimulate antibodies against HIV, as well as generating an effective CTL response. "I think you need both arms of the immune system," says John Moore, an AIDS researcher at the Weill Medical College of Cornell University in New York. However, Andrew McMichael, an immunologist at the University of Oxford, believes researchers should develop and test therapeutic vaccines that work on each arm of the immune system separately, so that each effect can be understood in isolation. If the two approaches are combined too early, he suggests, the results will be difficult to interpret.

One example of a candidate therapeutic vaccine that elicits antibody production as well as a CTL response is an inactivated form of an HIV protein called Tat. This protein contributes to the damage wrought by HIV in several ways: inside host cells, it directs the transcription of HIV genes; outside, it seems to hasten the death of uninfected CD4⁺ cells and stimulates them to produce more chemokine receptors, rendering them more susceptible to infection. Based on encouraging results from primate studies¹⁵, Robert Gallo, director of the Institute of Human Virology at the University of Maryland, believes that a vaccine based on the inactivat-

ed Tat, or Tat toxoid, should be included in a multicomponent therapeutic vaccine.

Pantaleo agrees that no single vaccine is likely to control HIV infection. He also points out that the most promising candidate vaccines being developed for prevention have not yet been tried as therapeutic vaccines. To devise an effective therapeutic strategy, he suggests, researchers may need to combine several of these newer vaccines.

Faced with these challenges, it seems that AIDS researchers are in for a long haul. But many refuse to be downcast. While the growing realization of the importance of viral reservoirs has dashed most researchers' hopes of eradicating HIV from the body, elimination may not be necessary to manage the disease. "If you could have a therapy that was easy, cheap, not toxic, not inconvenient, worked without side effects, could work life-long, but you didn't eradicate, is that OK?" asks Gallo. "Sure it's OK."

Paul Smaglik is Nature's Washington DC correspondent.

1. Wei, X. *et al.* *Nature* **373**, 117–122 (1995).
2. Ho, D. D. *et al.* *Nature* **373**, 123–126 (1995).
3. Chun, T.-W. *et al.* *Nature Med.* **1**, 1284–1290 (1995).
4. Wong, J. K. *et al.* *Science* **278**, 1291–1295 (1997).
5. Finzi, D. *et al.* *Science* **278**, 1295–1300 (1997).
6. Chun, T.-W. *et al.* *Proc. Natl Acad. Sci. USA* **94**, 13193–13197 (1997).
7. Chun, T.-W. *et al.* *Nature* **401**, 873–874 (1999).
8. Davey, R. T. *et al.* *Proc. Natl Acad. Sci. USA* **96**, 15109–15114 (1999).
9. Finzi, D. *et al.* *Nature Med.* **5**, 512–517 (1999).
10. Zhang, L. *et al.* *N. Engl. J. Med.* **340**, 1605–1613 (1999).
11. Ramratnam, B. *et al.* *Nature Med.* **5**, 82–85 (2000).
12. Ruiz, L. *et al.* *AIDS* **14**, 397–403 (2000).
13. Chun, T.-W. *et al.* *Nature Med.* **5**, 651–655 (1999).
14. Maino, V. C. *et al.* *AIDS Res. Hum. Retroviruses* **16**, 539–547 (2000).
15. Panza, C. D. *et al.* *Proc. Natl Acad. Sci. USA* **97**, 3515–3519 (2000).

AIDS dissidents aren't victims — but the people their ideas kill will be

Sir — As South African scientists working in the field of HIV/AIDS vaccine research, we are extremely concerned about the letter president Thabo Mbeki recently sent other heads of state (*Nature* 404, 911; 2000). As an individual Mr Mbeki is entitled to his point of view, but as our head of state we feel he risks binding our country to an untenable position.

Mr Mbeki's comments that distress us most are these:

1: "It is suggested ... that there are some scientists who are 'dangerous and discredited' with whom nobody ... should communicate or interact ... We are now being asked to do precisely the same thing that the racist apartheid tyranny we opposed did, because, it is said, there exists a scientific view that is supported by the majority, against which dissent is prohibited."

This is unfair. The views of the 'AIDS dissidents', publicly aired when the debate was current, are largely ignored now because most experts do not believe they have any currency in the light of today's knowledge. Yet the 33-member committee Mr Mbeki has convened to advise his government contains as many 'AIDS dissidents' from other countries as South African scientists. Less than half of the total are HIV/AIDS experts (see *Nature* 405, 105; 2000).

2: "The scientists we are supposed to put into scientific quarantine include Nobel prizewinners, members of academies of science and emeritus professors of various disciplines of medicine!"

This is a misleading statement: distinguished though these people may be, if they have not worked in areas concerning HIV/AIDS they may not be well-enough informed to have credibility in this debate.

3: "People who otherwise would fight very hard to defend the critically important rights of freedom of thought and speech occupy, with regard to the HIV/AIDS issue, the frontline in the campaign of intellectual intimidation and terrorism which argues that the only freedom we have is to agree with what they decree to be established scientific truths."

This is incorrect. 'AIDS dissidents' promote the idea that unholy alliances of pharmaceutical companies and funding bodies are bent on silencing them. The fact is that, if one's scientific views are very obviously not being backed up by other people's findings, one's scientific credibility is lessened. Internationally, science is a democratic

institution: as such we would hope that Mr Mbeki would sympathize with it. This case has clear historical parallels with the championing of Trofim Lysenko's flawed science by the authorities in the former Soviet Union, and with the 'scientific' justifications of apartheid by the old South Africa. Neither is a good example to follow!

4: "It may be that these comments are extravagant. If they are, it is because in the very recent past, we had to fix our own eyes on the very face of tyranny."

This is irrelevant to the country's AIDS-related crisis. The previous government was guilty of inaction in the face of a threatened epidemic; the present government has not done enough in the past five years to stave off the disaster that now threatens us.

We would like Mr Mbeki and others to consider how the mass of South Africans would react if he were to give a sympathetic ear to unrepentant proponents of apartheid. His willingness to be influenced by people with no credibility causes as much anguish to those of us working to combat HIV/AIDS.

The simple facts, as shown by a huge volume of scientific and medical research, are that HIV causes AIDS; that in Africa (as in other developing regions) the disease is mainly spread heterosexually; and that AIDS kills poor people in disproportionate numbers. We most emphatically do not need to revisit the debate on the causation of AIDS. What we do urgently need is to educate, train and medicate, to save lives.

As long as Mr Mbeki is being advised by people with no credibility, we as South African scientists feel dangerously marginalized in the search for solutions to HIV/AIDS. **Edward Rybicki***, **Anna-Lise Williamson†**, **Lynn Morris‡**

*Department of Microbiology, University of Cape Town, PB Rondebosch 7701, South Africa

†Department of Medical Microbiology, University of Cape Town, PO Observatory 7925, South Africa

‡National Institute for Virology, Private Bag X4, Sandringham, Johannesburg 2131, South Africa
The authors are members of a consortium sponsored by the South African AIDS Vaccine Initiative, seeking to formulate vaccines to the HIV-1 subtype C viruses prevalent in the area.

Bureaucracy strangles Latin American research

Sir — In his special supplement on science in Latin America last year, Colin Macilwain¹ rightly pointed out that scientists in most countries in the region complain about excessive bureaucracy and the time wasted by customs authorities in processing scientific equipment and materials for research.

But the major problem is in the enormous, inefficient bureaucratic system that reigns in almost all Latin American government institutions, not just in universities.

The system does not serve academic researchers: it creates all types of administrative bottlenecks. Most universities employ more administrative staff than teaching staff, using up a major part of the budget in unnecessary paperwork as well as for their salaries and bonuses.

The general administration of a university maintains four or five divisions with titles such as "purchasing", "budget", "finance", "control" and so on, each section maintaining large numbers of staff, each with a specific role. If one person does not function, the rest stand still. The simplest transaction, for example payment of expenses to attend a scientific meeting, will probably not be completed until the event is over, unless the researcher personally tracks the file from desk to desk. Every four years, the authorities change and the new ones introduce their own systems, allegedly for efficiency but in fact to justify hiring new people. So a new bottleneck is created. This growth of administration can come only at the expense of research and teaching².

The causes of administrative growth in Latin American universities are complex. This growth probably stems partly from Cordoba's university reform³ in 1918, and partly from earlier movements for the democratization of the Latin American universities in Uruguay, where the idea of an autonomous democratic university first emerged. Cordoba's reform movement was an important influence in replacing the autocracy of Latin American universities with a democratic campus system. But later developments — including an excess of politics and the development of trades unions among the administrative staff and students — have eroded the original initiative in many universities.

Drastic measures are required to minimize bureaucracy and improve efficiency, but university authorities (which are elected) will not risk their popularity. A few countries in the region have had some success in cutting their administrations down to size, but this has not been possible in others because of powerful unions and rigid control by the administrative system.

University managements have a responsibility to create a more dynamic, efficient research environment. Governments in the region should hold them responsible for unproductiveness in research and teaching.

Julio E. Pérez, **Abul K. Bashirullah**

Instituto Oceanográfico, Universidad de Oriente, Apartado Postal 138, Cumana 6101, Venezuela

1. Macilwain, C. *Nature* 398 supp., A10 (1999).

2. Finn, R. *The Scientist* 11, 13 (1997).

3. Penálver, L. M. *Teoría y Praxis de la Universidad a Distancia*. (Fondo Editorial para el Desarrollo del Educación Superior, Caracas, 1992).

Emblem of a golden era

A physiologist's role in the coming of age of American biomedical science.

Walter B. Cannon: Science and Society

by Elin L. Wolfe, A. Clifford Barger & Saul Benison
Harvard University Press: 2000. 630 pp.
 \$30, £18.50

W. F. Bynum

Most biographers and readers of the genre will know that the second half of a person's life is often less interesting than the first. (I write this from the second.) It is as if Aristotle were more prescient than Plato, and becoming is exciting, while simply being is rather dull.

Walter Cannon's life may be an exception to this rule. For one thing, he began the second half with the war to end all wars; for another, he managed until the last few years of his life to keep an active research agenda going, even while his sense of civic duty kept the 'society' part of his activities at full throttle. Consequently, the second of this brace of volumes on the eminent American physiologist maintains the standard of the first, which appeared more than a decade ago (*Walter B. Cannon: The Life and Times of a Young Scientist*, Harvard University Press, 1987). The intervening years have taken their toll, with the death of Clifford Barger and the serious illness of Saul Benison. It has thus been left to Elin Wolfe to complete the monumental task that the trio began. The project was made more challenging, but also more rewarding, by the extensive archival collections relating to Cannon and his colleagues in repositories at Harvard University and other American and British institutions.

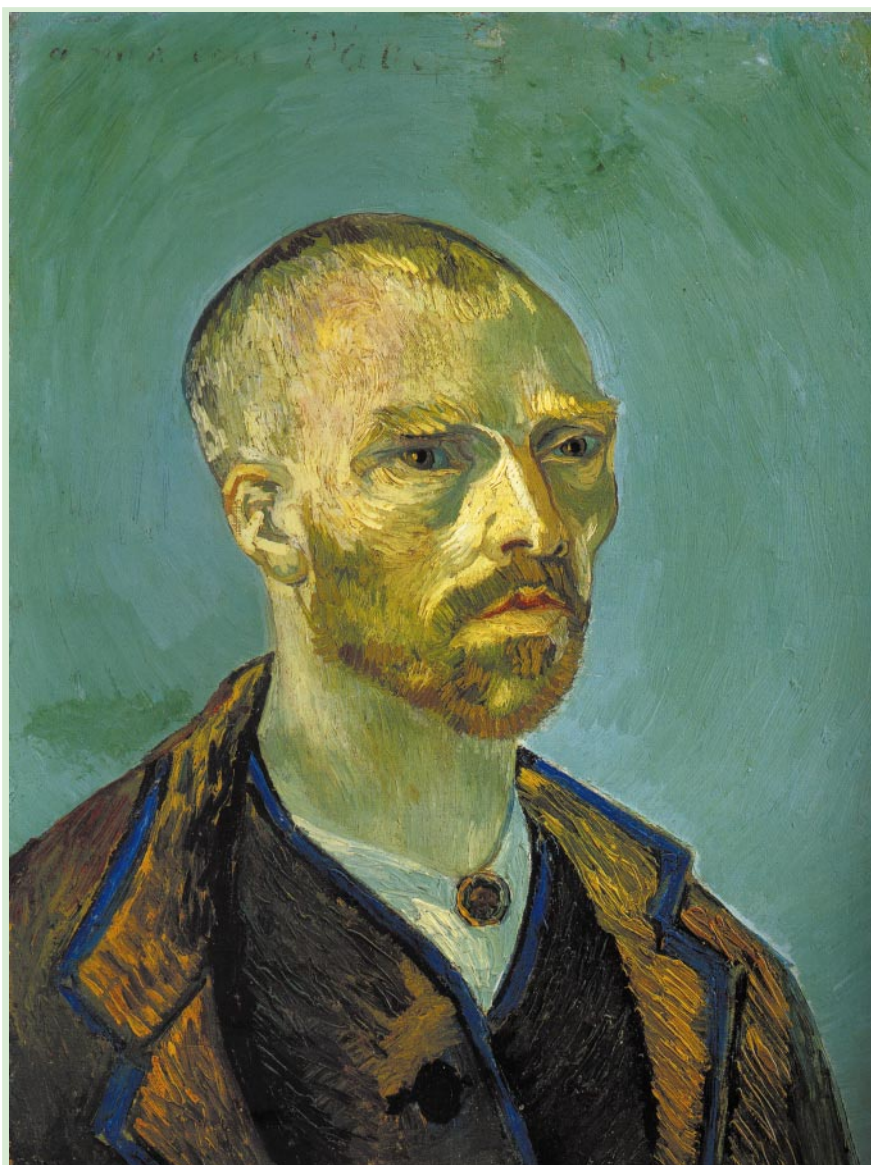
Cannon lived through the golden era of whole-animal physiology. Along with Ernest Starling and Sir William Bayliss, he was an experimentalist of such scope and originality that his absence from the Nobel prize lists seems surprising. He might have been considered when Sir Henry Dale and Otto Loewi shared the 1936 prize for their work on neurohumoral transmission at nerve synapses, although his mature theory of what actually happens at the synapse received little contemporary assent and did not stand the test of time. His researches on the physiological bases of emotional expression brought him to the attention of clinicians.

Of lasting quality was Cannon's work on regulatory mechanisms, embodied in his continuing preoccupation with homeostasis (a word he coined). Cannon's *Wisdom of the Body* (1932; out of print), one of a handful of modern physiological classics, is probably unique in its presentation of an important and novel synthesis in language that could be

appreciated by a non-specialist audience. It thoroughly deserved its popular acclaim. The last decade of his life was so clouded by ill health that he was driven wryly to remark that maybe the body was not so wise after all.

Cannon was already a highly visible scientist by the 1930s, but he also liked to describe himself as a hermit in the laboratory, even if his public as well as his domestic life remained exceptionally full. He was ever

an espouser of good causes. He served for years as the chairman of the American Medical Association's Committee for the Protection of Medical Research, a position that brought him face to face with the antivivisection movement. He was deeply touched by the Spanish Civil War, partly because of his Spanish students and colleagues, and devoted many hours to raising money for the relief of Republican groups. Visits to Russia



Expressions that stand the test of time

Vincent van Gogh wanted to create portraits that would "appear as revelations to people in a hundred years' time". He believed them to be a means of showing "the soul of the model". *Self-Portrait*

Dedicated to Paul Gauguin, shown here, was painted in 1888, two years before van Gogh's death. (From *Van Gogh Face to Face: The Portraits* by Roland Dorn et al.; Thames & Hudson, £32, \$50.)

and China inevitably involved him in the politics of communism and fascism. His ties with Russia were strengthened by his admiration for, and friendship with, I. P. Pavlov, whose well-being was the object of so much concern among the international physiological community.

Above all, Cannon had faith, as both a scientist and a citizen, in the capacity of science to transcend petty nationalism, and this faith was not broken even as he watched the abuse of science and rationality by Nazi spokesmen. He was concerned with the plight of Jewish refugees, although the faculty at Harvard was dominated by male Protestants who mostly preferred to aid their Jewish colleagues at a distance.

Harvard was, in the fullest sense, Cannon's Alma Mater. He was an undergraduate and medical student there, and spent his professional career guiding its department of physiology to a position of international pre-eminence. He was eventually able to secure recurrent funding from the Rockefeller Foundation, which enabled him to nurture his department through a period of sustained expansion. He worked with a succession of graduate students and fellows from all over the world, including Asia and South America. By the time he died, in the closing days of the Second World War, his students occupied chairs in key universities throughout the world. His had been a consistent voice, at Harvard and wherever he went, arguing for the value of medical research for human welfare, and the fundamental importance of free enquiry as a basic necessity for democratic societies.

Cannon's breadth of interests and activities endows his career with an emblematic quality. This volume is consequently as much about the maturation of American biomedical science as about the life of one individual. It is about Big Science in the making.

At the same time, Cannon's world seems incredibly remote from our own. Even at the time of his retirement, his department was minuscule by modern standards, with only a couple of other permanent members on the teaching staff. A few thousand dollars a year from the Rockefeller coffers made all the difference. Recruitment of students and fellows was often by word of mouth and can seem to us rather casual. Cannon published many research papers alone and rarely with more than one or two collaborators. Research assessment exercises were few and far between. He actually ran his own department.

Before anyone gets too nostalgic, however, we should remember that some features of modernity were also part of Cannon's scene. Many of his junior staff faced years of hand-to-mouth uncertainty. Jobs were scarce, especially during the 1930s, and Cannon spent his share of time fretting

about what would happen to his pupils. Experiments did not always go well, and priority disputes were not unknown. There was always too much to do.

On balance, though, these two substantial volumes devoted to this Harvard physiologist suggest that the first half of the twentieth century was not a bad time to be Walter Bradford Cannon. ■

W. F. Bynum is at the Wellcome Institute for the History of Medicine, 183 Euston Road, London NW1 2BE, UK.

Witness after the event

A Fly for the Prosecution: How Insect Evidence Helps Solve Crimes

by M. Lee Goff

Harvard University Press: 2000. 240 pp.

\$22.95, £14.50

Mark Benecke

"I didn't think anyone would actually be interested in reading about my cases," wrote Lee Goff about *A Fly for the Prosecution*, his first popular-science book. Such a lack of interest, however, was unlikely, as Goff is one of a small number of scientists who work out how long a corpse has been dead by using maggots and other creepy crime-scene reconstruction assistants. Furthermore, his lectures and workshops are constantly overcrowded and his sense of humour, combined with scientific accuracy, makes him the darling of his various audiences.

Although forensic entomology has a certain fascination for the public, any forensic pathologist will tell you that it is much more entertaining to read about cases than to collect live arthropods from decaying human corpses. This is where Goff's book comes in — it is a colourful collection of forensic entomology research and cases (mostly his own in Hawaii), along with personal thoughts about how to deal with violent death and decay, and the story of his life.

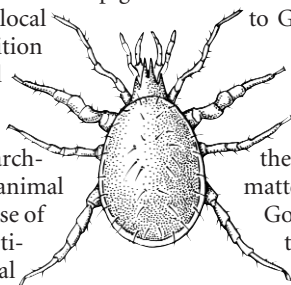
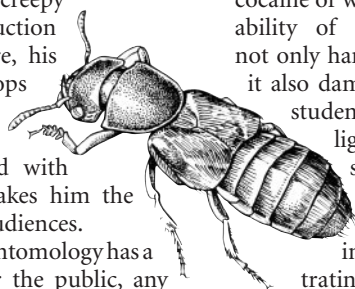
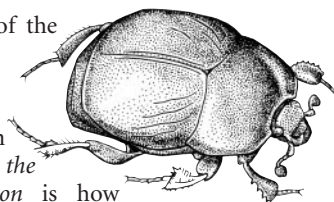
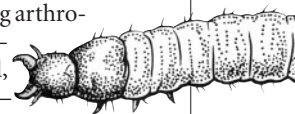
An expert on mites, Goff stumbled into forensic entomology after hearing a talk about the decomposition patterns of pigs in 1981. He quit his job at a local museum and accepted a position at the College of Tropical Agriculture in Hawaii. Soon after, he and a handful of colleagues, who were either researching the activity of insects on animal cadavers or looking into the use of entomology in criminal investigations, set up an informal

group called the Council of American Forensic Entomologists (CAFE). It is highly enjoyable to follow motorcycling arthropod specialist Lee Goff — at that time, with his beard, earring and shaggy hair — to formal forensic academy meetings, into the FBI Academy, in front of highly offensive US attorneys in murder trials, and away from the preparations of a peaceful New Year's Eve dinner to collect evidence at a crime scene. Thanks to the work of CAFE members, the method has grown in popularity among US forensic pathologists, forensic scientists and the police, until now the FBI Academy includes an annual arthropod collection class in their course "Recovery of Decomposed Bodies". The CAFE-edited book *Entomology and Death: A Procedural Guide* (Joyce's Print Shop, 1990) was a collection of practical information on the subject.

One of the most valuable points made in *A Fly for the Prosecution* is how different the characteristics of insect populations are in contrasting habitats such as corpses that are wrapped or burned, those hanging above ground compared with buried ones, and those poisoned with cocaine or with other drugs. This high variability of most biological processes is not only hard to explain in front of a jury, it also dampens the enthusiasm of many students and beginners. Goff sheds light on the limits of the method, stressing that forensic entomology needs experienced practitioners. As he points out, any investigation starts to get frustrating the moment identification of an uncommon arthropod species has to meet a deadline set by a subpoena instead of by common sense.

It can be difficult to persuade research agencies of the need for additional scientific studies, for example on the development of local arthropods at different temperature levels, on the identification of toxins produced by insects that have fed on corpses, and on DNA typing of insects. Therefore, it has taken years to bring the method into routine use in forensics. It is thanks to Goff and his US colleagues that forensic entomology is now widely known as an adequate and highly effective tool in difficult questions concerning the time of death and related matters of criminal investigation. Goff's book is the culmination of this effort.

The method was discovered

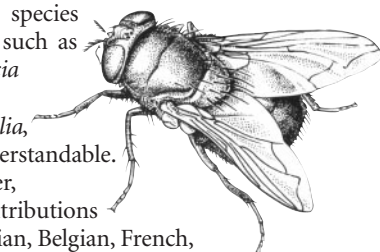


and pioneered in Central Europe 150 years ago. In Europe and Asia, where the older generation tends to view any popularization

of science as suspicious and unnecessary, forensic entomology was kept alive after the war by only four scientists. Such historical details are only briefly mentioned by Goff. For the same reason, the use of non-metric units and Americanized species names, such as *Phaenicia* instead of *Lucilia*, is understandable. However,

the contributions of Russian, Belgian, French, Finnish, Canadian and German researchers should not be underestimated. Therefore, and to get additional perspectives into forensic entomology, I would recommend three further recent publications. One is a special issue of the journal *Forensic Science International* on the subject (of which I am the guest editor), which will also include the latest research results to be presented at an international meeting of forensic entomologists held in Brazil in August 2000 (<http://www.benecke.com/feauthor.html>). There is also a well-researched popular-science book by Jessica Snyder-Sachs about to be published, and the scientific handbook *Entomological Evidence: The Utility of Arthropods in Legal Investigations* (CRC Press, 2000). Meanwhile, enjoy Goff's beautifully illustrated, scientifically sound and entertaining story. ■

Mark Benecke is in the Forensic Biology Research Unit, Institute for Zoology, University of Cologne, Cologne 50923, Germany.



Of viruses and men

Virus: The Co-discoverer of HIV Tracks its Rampage and Charts the Future

by Luc Montagnier
W. W. Norton: 1999. 256 pp. \$24.95, £18.95

Robin A. Weiss

The book telling the story of AIDS should be a thriller. The explosion of a deadly novel disease involving sex and drugs, its origins and the ensuing pandemic, the unfolding epic of finding its cause, the rivalry involved, the rapid development of science-based diagnosis and later treatment, how Western gays empowered patients, the hapless scene in Africa—all these facets could make gripping reading. We await an inspired writer to take up the challenge.

When Bob Gallo's *Virus Hunting* (Basic Books, 1991) was published, Mirko Grmek,

the author of *History of AIDS* (Princeton University Press, 1993) who sadly died on 6 March, remarked that "Generals write memoirs; history is best left to historians". In the battle for credit for the discovery of HIV, the general leading the French troops followed with his own memoirs, *Des Virus et des Hommes* (Odile Jacob, 1994), now updated and translated under the title *Virus*. Luc Montagnier's book comprises a brief account of his career, the discovery of HIV, and chapters dealing in lay terms with HIV and AIDS as a global problem.

While Montagnier's experiences as a young scientist in Britain are told with wry humour, most readers will zoom in on his account of the race to isolate and characterize HIV. This is by far the most interesting section of the book, whereas the second half, about the understanding, treatment and proposed prevention of HIV infection, seems rather mundane.

Unlike contemporaneous diaries, memoirs benefit from hindsight. Naturally, the author of *Virus* portrays himself in a favourable light, although he is no more self-serving than one reviewer of his book. Besides, although the French were treated shabbily over the discovery of HIV, Montagnier won this battle, as the record of research papers makes clear.

The Institut Pasteur team published their preliminary observations on a new human retrovirus linked to AIDS in *Science* in May 1983, and by April 1984 had published further papers on new isolates from AIDS patients and had found that HIV was more closely related to animal lentiviruses (by its morphology and cell-damaging effect) than to the human T-cell leukaemia viruses (HTLVs). The American team first presented their evidence for an AIDS virus in May 1984 and throughout that year maintained that it was related to the HTLVs. When the gene sequences were published in January 1985 (*Nature*, Montagnier notes, declined to consider the French paper), it became apparent how distinct HIV was. And when a West African man in Portugal developed AIDS without evidence of HIV-1 infection, Montagnier's group identified the second AIDS virus, HIV-2.

For a victorious general renowned throughout the world, Montagnier remains surprisingly bitter. He portrays himself as a prescient though misunderstood scientist who, almost alone, saw the danger of the AIDS agent contaminating the blood supply but was ignored. He complains that a plea to expand his laboratory went initially unheeded by the Institut Pasteur, believing that extra funds should have flowed automatically from his letter—perhaps he has never needed to write a proper research grant application. He feels that his theory that progression from HIV infection to AIDS requires an essential cofactor to

New in paperback

Eco-pragmatism: Making Sensible Environmental Decisions in an Uncertain World

by Daniel A. Farber

University of Chicago Press, \$15, £10.50

"Daniel Farber's is a timely and well-argued contribution to the understanding of how judicial systems can maintain the balance between competing environmental and economical claims. It does more than that: it puts forward a framework for using judicial pragmatism to integrate environmental considerations into economic activities." Calestous Juma, *Nature* 399, 653–654 (1999)

The Descent of Mind: Psychological Perspectives on Hominid Evolution

edited by Michael C. Corballis & Stephen E. G. Lea

Oxford University Press, £17.99

Green Chemistry: Theory and Practice

by Paul T. Anastas & John C. Warner

University of Chicago Press, \$25, £17.50

"... as Anastas and Warner point out in *Green Chemistry: Theory and Practice*, synthetic chemists have not been particularly environmentally conscious... The book defines the 12 principles of green chemistry—described as the Hippocratic oath for chemists... Anastas and Warner's prediction of future trends is a mixed bag of well-defined goals." Roger Sheldon, *Nature* 399, 33 (1999)

Intellectual Impostures

by Alan Sokal & Jean Bricmont

Profile Books, £6.99

The Nature of the Book: Print and Knowledge in the Making

by Adrian Johns

University of Chicago Press, \$22.50, £16

"The publication of science books, as much as any kind of book, is a complex process, involving a range of parties and interests, not all with common aims. It is basically this insight, applied to the burgeoning scientific culture of seventeenth-century England, that underlies Adrian Johns's provocative and stimulating, if ultimately not wholly satisfactory, book." Michael Hunter, *Nature* 396, 534–536 (1998)

Virtual Organisms: The Startling World of Artificial Life

by Mark Ward

Pan, £6.99

The Great Disruption: Human Nature and the Reconstitution of Social Order

by Francis Fukuyama

Profile Books, £8.99

HIV met too cool a response from the scientific community, although he has softened his fixation on *Mycoplasma* as the essential cofactor in this new English edition.

Montagnier depicts his style of leadership as authoritarian rather than persuasive. He "instructs" fellow scientist Jean-Claude Chermann to set up serological assays, and he "assigned to Marc Alizon the arduous task of cloning" HIV, to be sequenced with Simon Wain-Hobson and Pierre Sonigo. I think Montagnier does himself injustice here, for, by the time I arrived on sabbatical at the Bâtiment du SIDA, the *pasteuriens* were living up to Emile Roux's ideal "where individuals could contribute to a common goal without risking their intellectual freedom". But by then, Chermann had left for a sunnier clime, while Alizon and Sonigo had moved to the Institut Cochin.

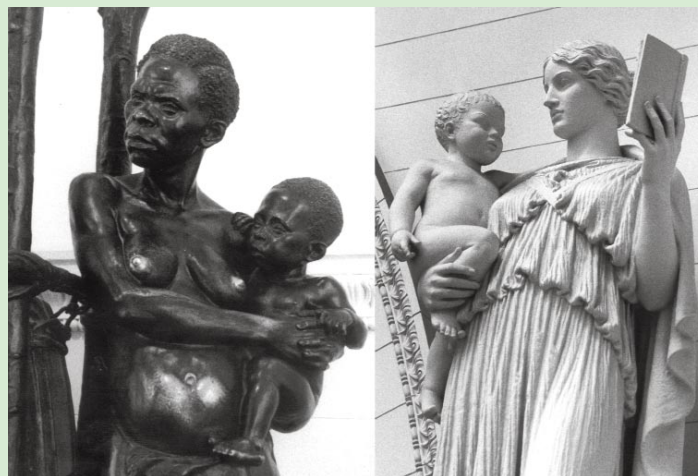
Montagnier claims he gained his insight into AIDS because he trained in medicine as well as science. "Clinicians take care of the sick. Researchers take care of lab work. And most importantly, no one can stray from his or her corner. I have taken the risk of transgressing these customs." But after a generation of MD/PhD programmes, Montagnier is by no means unique. He wrongly states that "Gallo was not a medical doctor, but rather a biochemist by training", and that this led Gallo to "misunderstandings and blunders". This general does not know his adversary.

In the epilogue, Montagnier explains how he is using his fame to fight the real enemy, AIDS. "I have realised that my position as a pioneer researcher and the notoriety I have achieved should help me to reach beyond my research activities and contribute to resolving the problem of AIDS worldwide." Accordingly, he set about establishing three centres, in a private hospital in Paris, in Abidjan on the Ivory Coast, and at Queens College, New York, with further centres promised across the world. Alas, not long after writing of this grandiose enterprise, more bitter pills are having to be swallowed as the New York and Paris schemes collapse. As with biotechnology companies, pioneers do not always make the best chief executive officers.

Overall, the book is illuminating not so much about AIDS as about the attitudes and vicissitudes of one of its major players. Now that the two opposing generals have set down their memoirs, the troops will hope they can march forward peacefully towards better treatment and an effective vaccine. There is, however, one further memoir I should like to read but doubt will be written — that of Françoise Barré-Sinoussi, the Rosalind Franklin of HIV. ■

Robin A. Weiss is at the Windeyer Institute, University College London, 46 Cleveland Street, London W1P 6DB, UK.

Science in culture



Ethnic meets aesthetic: visitors to the museum can view Malvina Hoffman's *Pygmy Family* and Henry Herring's *Dissemination of Knowledge* together.

Type and archetype Hoffman and Herring in the Field Museum, Chicago.

Martin Kemp

The Field Museum in Chicago, built to house the biological and anthropological collections for the World's Columbian Exposition in 1893, is one of those remarkable foundations in which the creations of nature and 'artistic masterpieces' stand cheek-by-jowl as much by accident as by design. Devoted to natural history and anthropology, those fraternal twins of nineteenth-century science, the museum contains a rich treasury of artefacts originally collected from exotic cultures as items of ethnographic significance. Many, such as African masks and oriental jade, would now seem equally at home as aesthetic treasures in the nearby Art Institute.

A few works were ordered specifically from contemporary artists as a more conscious effort to bring 'art' into the museum. In 1930, the notable sculptor Malvina Hoffman was commissioned to create a series of bronzes of ethnic types from around the world. The commission had a distinguished, mainly French ancestry, most notably the striking ethnic busts by Charles-Henri-Joseph Cordier assembled from 1851 onwards in the anthropological gallery of the Muséum National d'Histoire Naturelle.

Born in New York in 1885 and trained by the great Auguste Rodin in Paris, Hoffman had gained a reputation for strongly characterized portraits and family groups. For the Field project she produced not only busts but also life-size figures, often in action, and ingeniously staged groups. The *Pygmy Family* from the Ituri Forest in northeast Congo, sculpted in 1931, gives a good idea of the powerful modelling and vivid portrayals characteristic of the 104 works that more than adequately met the museum's expectations.

Almost three-quarters of a century later, the conceptual framework within which we approach the works has become more complex. No longer displayed in the Hall of the Races of Mankind, they are now dispersed in perambulatory spaces.

The 'politically correct' curator of today is likely to feel a tinge of unease in the face of such picturesque portrayals of exotic racial types as objects of artistic curiosity. Immediately at the top of the stairs, a display featuring busts of a Zulu and a Padang woman (the latter's neck extended by stacked rings) accordingly invites us to consider the cultural construction of beauty. Photographs of Elisabeth Vigée-Lebrun's elegant portrait of Marie Antoinette, Rubens' voluptuous *Feast of Venus* and bikini-clad sunbathers on a Californian beach aspire to render Western conventions as curious as the strange customs of African 'natives'.

The museum's free brochure-map seems to signal unease when it exhorts each visitor to the "Exhibitions about Culture" to consider how "some of these exhibits may portray people or ideas that are new to you. Thank you for viewing all the displays with respect for the traditions and environments they represent."

One potentially telling element in the museum's visual ensemble of artistry is now more likely to be overlooked than treated with active disrespect. In the main hall, high above Carl Akeley's compelling taxidermic drama of two elephants fighting, stand two immaculately white statues representing *Science* and *The Dissemination of Knowledge*, commissioned from Henry Herring in 1915. Standing on the balcony, we can stare directly across from the lumpily uncompromising portrayal of the bronze woman and child in Hoffman's *Pygmy Family* to the Grecian perfection of Herring's high-minded allegory.

As with so much anthropological characterization, the aesthetic ideal set by ancient marble Venuses and the *Apollo Belvedere* continued to provide the ingrained visual filters through which other works were evaluated. Ethnic type stood characterized by contrast to aesthetic archetype. Is this a habit we have wholly discarded? ■

Martin Kemp is in the Department of the History of Art, University of Oxford, 35 Beaumont Street, Oxford OX1 2PG, UK.

More is less

Economists and governments lag decades behind Derek Price's thinking.

Terence Kealey

"Eighty to 90 per cent of all the scientists who have ever lived are alive now". This statement has been true for every year since 1700, perhaps even earlier. Moreover, "any scientist, looking back at the end of his career, will find that 80 to 90 per cent of the scientific work [in his field] has taken place before his very eyes". Every retired scientist is a walking, living, eye-witnessing historian of most of the science that has moulded his or her discipline.

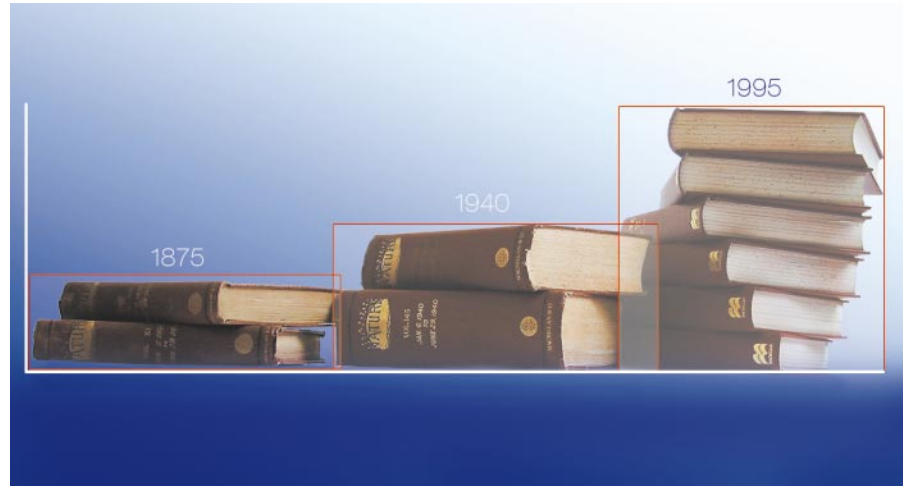
This is how Derek de Solla Price opened his book *Little Science, Big Science* (Columbia University Press, 1963). These aphorisms are well known, but few remember the man who coined them, nor how he discovered them, nor how well he wrote about them. Even fewer people have internalized their significance.

Price was an ex-physicist who decided to "turn the tools of science on science itself", and his findings were memorable. He plotted, for instance, a graph showing the numbers of scientific journals founded since the first two — the *Philosophical Transactions of the Royal Society* and the *Journal des Sçavans* — were launched in 1665. This showed that science's exponential growth over the last three centuries has been astonishingly steady.

Although the graph does not adjust for the journals that fail, the doubling time since 1700 in the numbers of journals has been around 15 years and, because journals expand in size, the doubling time in the numbers of papers has been about 10 years. And, as far as Price could chronicle, the doubling time in the numbers of scientists and engineers has, since 1700, been between 10 and 15 years. So we see that, depending on the parameter measured, the entity we call science has, over the past three centuries, doubled in size every 10 to 15 years.

How different from non-scientific history! Most historical events took place long before any of the readers of this article lived, and, although human populations are growing, the majority of people who have lived will always be dead.

In other graphs, Price showed that many other measurable aspects of science also show tightly predictable rates of exponential growth. For example, the numbers of papers published annually with different multiples of authors can be extrapolated into the future, as can the asymmetric distribution of the numbers of papers published annually per author (Lotka's law), the numbers of



Bigger and better? The quantity of material published in *Nature* (not to mention its growing number of 'sister' journals) mirrors the exponential growth of all areas of science.

universities founded annually, and the likelihood of any discovery being made independently.

We scientists are, these days, no longer much excited by such data because we have grown only too familiar with citation indices, impact factors and other quantitative tools of research harassment, but Price, on whose work those academic punishments are largely based, despised them as "counting nonsense". Price saw his work as a contribution to economics — but there he was to be disappointed, for the economists have ignored him.

And the loss has been theirs. There is an extraordinary school of economics currently rampant called 'endogenous growth theory' which is based on a series of beliefs about science that Price exposed as errors a third of a century ago. Such errors might not matter, except that the US and UK governments, to name but two, have based their policies for long-term economic growth on the theory.

The theory states that economic growth depends on scientific and technological advances — which is true — yet it also

proposes that science, miraculously, is like a perpetual motion machine, in that if you know twice as much science you grow four times as rich (so you can afford four times as much science, and so on). But, as Price showed, science demonstrates diminishing returns. The rate of scientific growth is about twice that of economic growth, which means that you have to do four times as much science to get twice as rich. One day, as science's exponential demands on national incomes become excessive, the rates of scientific — and therefore economic growth — will slow.

Moreover, endogenous growth theory states that, because science is universally available, it must be funded by the state. Well, science may need state funding, but not because it is universally available. As Price showed, science is organized in "invisible colleges", each consisting of a microdiscipline of a few hundred researchers who understand the field and its tacit, unpublished lore; but people outside the invisible colleges are disenfranchised, and a paper in high-energy physics is almost as obscure to a biochemist as to a historian.

Finally, the endogenous growth theorists are excited by their discovery that skilled people emigrate from poor countries to rich ones — something Price described as "the brain drain" decades ago.

There are few more important topics than the economics of science, but we scientists have erred in entrusting it to economists. By ignoring Price, they have incubated 37 years of mistakes. ■

Terence Kealey is in the Department of Clinical Biochemistry, University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 2QR, UK.

Science grows at twice the rate of the economy — so you have to do four times as much science to get twice as rich.

Consequences

Perhaps the biotech sector should have employed more philosophers.

Poul Anderson

The United States is merely the first country to confront the question. Soon others must. It might have been foreseen, but was not, any more than was the explosive growth of the Internet in the last century. You young people had better understand the situation. You will be coping with the results of whatever decision we make today. Let me tell you in simple words how it all came about. This ought to be common knowledge, but seems nearly lost in the current furore.

As usual, many factors operated together. The principal ones here, I think, were the synergistic advances of computer science, biochemistry, and nanotechnology. Even such bare beginnings as the sequencing of DNA would obviously have been impossible without great computer capabilities. Likewise for work at the molecular and atomic levels, such as the creation of assemblers and nanomachines. It seemed to most enthusiasts that the future, including the rise of artificial intelligence to consciousness, lay with silicon.

Yet carbon, mainly in messy, gooey proteins, had been nature's choice for nanotechnology. Several billion years of success spoke for it. Early experiments with biological computers held promise. However, rapid and stupendous progress in electronics and, presently, photonics overshadowed them.

Then it turned out that its curve was not exponential, but asymptotic. Although we seemed nowhere near the limit of data-processing capability, it was only as good as its programming. A machine could certainly do much of this, too. In the end, though, the business came back to humans. If nothing else, they alone could determine what purpose a program should serve — what it was for. No matter how awesome, inorganic computers lacked creativity, imagination.

Reluctantly, workers in the field were forced to admit that the thoughts of Roger Penrose in the twentieth century had been right. The conscious, originating mind is not totally algorithmic. Of course, this did not imply mysticism or vitalism. It meant just that the proper model was the human brain. Nature had, so to speak, known what she was doing when she went in for carbon.

Thanks largely to silicon, nanotechnology and biochemical knowledge had reached the point of biosynthesis. The technology was, indeed, already routine in many applications, such as the design and production of

microbes to destroy pathogens and malfunctioning cells. Making a brain to order would be a long step forward, but possible. It would be integrated with a computer, combining their functions to form an intellect that might prove half-godlike.

At first the results of the effort were disappointing. Again workers found themselves harking back to the twentieth century, in this case to the neurologist Antonio Damasio. He had explained that the human brain is not a calculator in the skull, it is part of an organism and, in fact, itself a gland. Emotions, motivation and inspiration do not spring from reason, but they are vital to it if it is to be effective. Furthermore, sanity requires constant sensory input, not merely through electrophotonic channels but through a whole body interacting with the whole world.

Thus it was necessary to develop a complete organism. We knew it could be simpler than we are. Evolution built on whatever happened to be there, yielding systems that were often needlessly complicated. For instance, most of the trace elements required in our diet were once toxins sequestered by the body. At the same time, the basic human animal is well engineered to experience reality and work upon it.

As you know, what we arrived at was the Linked Electronic–Organic systems, with radio relay connections to let the rather anthropomorphic creatures move freely about. LEOs do not, cannot, leap into existence fully knowing. Like us, they must have time to mature and learn. Also like us, they cannot do so without companions of their own kind. Hence today there are quite a few LEOs among us.

It goes without saying that our laws against cruelty always applied to them. Soon we could do no other than declare them fully human, with the same rights we enjoy. They reward us with their ideas and insights, which enrich us in many ways and which sometimes we more or less understand.

Now the LEOs want citizenship. After all, they point out, these days humans are conceived and gestated by a variety of processes. Why should theirs be considered special?

This looks reasonable, yes. But lately one of them has announced an intention to run for president. It would doubtless win.

What shall we do?

Or do we really have a choice?

Poul Anderson's novels include *Operation Luna*, *Harvest the Fire* and *Starfarers*, all of which are published by the Saint Martin's Press.



Recounting a genetic story

Roger H. Reeves

The DNA sequence of human chromosome 21, now published, provides indications that the total number of human genes has been overestimated, and is a valuable resource for research into Down syndrome.

The sequencing of whole genomes means that genetic information is increasingly considered in functional rather than structural terms. But for the moment, the principal structural element of genomes — the chromosome — remains the predominant unit by which to measure progress in the Human Genome Project.

On page 311 of this issue¹, a multinational consortium reports the complete sequence of chromosome 21, the smallest human chromosome. The results indicate that previous estimates of the total number of human genes may need to be revised downwards. Meanwhile, the small number of genes, and a catalogue that identifies them, provide a boost for those endeavouring to define all of the primary molecular players in Down syndrome. Affecting one in 700 live births, Down syndrome occurs when three copies of chromosome 21 are inherited instead of two (Fig 1, top). The condition is the most common known genetic cause of mental retardation and the leading cause of congenital heart disease, and results in a wide variety of other developmental and health problems.

The consortium's report¹ describes several new technical achievements. The total length of the sequence reported is 33.55 million base pairs (or megabases, Mb). This covers 99.7% of the long arm of chromosome 21 (Fig 1, bottom), and just exceeds the 33.46 Mb reported for the slightly larger long arm of chromosome 22 (ref. 2). The paper includes the longest continuous DNA sequence reported to date, extending 28.5 Mb. The entire chromosome sequence has only three gaps (totalling 100 kilobases), compared with the ten gaps (totalling about 1 Mb) for the long arm of chromosome 22.

The new sequence also includes 281 kilobases from chromosome 21's short arm — mapping and cloning of which posed a challenge because it contains several classes of highly repetitive sequences³. The length of this short arm can vary greatly among individuals. So this sequence is the first example of a large genome region that can expand or contract on a scale of many megabases.

The sequencing of the long arm of chromosome 21 provides a somewhat arbitrary, but nonetheless worthwhile, basis for deriving conclusions about the general organiza-



Figure 1 Chromosome 21 in context. Top, triplication of chromosome 21 is the genetic defect underlying Down syndrome. Bottom, transmission electron micrograph of chromosome 21, showing the long and short arms.

tion of the human genome. The most striking difference is the reduced gene content of chromosome 21 — 225 genes identified, compared with 545 on chromosome 22. The two consortia responsible for these sequences used somewhat different criteria to identify the genes within their respective

chromosomes (Box 1, overleaf). But the differences may well balance each other out, meaning that a comparison of gene numbers is valid.

Chromosome 21 was expected to be relatively gene-poor, but it seems that it is even more impoverished than anticipated. The

long arm of chromosome 21 represents about 1% of the human genome, but was predicted to contain less than 1% of the total number of human genes. The Unigene project⁴ suggested that chromosome 21 would contain only 80% of the number of genes that would be expected on the basis of its size. If the total number of human genes were 100,000, as predicted, chromosome 21 would still be expected to contain 800–1,000 genes. The 225 genes now identified¹ stand in stark contrast to this prediction.

Combining data from the long arms of the two completely sequenced chromosomes, the chromosome 21 consortium estimates that the human genome may contain as few as 40,000 genes. However, this is based on complete sequences for just 2% of the human genome, and could be low for a variety of reasons. For example, other human chromosomes may be more gene-rich. The major histocompatibility complex (MHC) region on chromosome 6 — a region essential to the immune system — spans only 3.6 Mb, but contains 128 genes and 96 pseudo-genes⁵.

Another measure of gene richness is provided by the number of 'CpG islands' on the long arms of chromosomes 21 and 22. These islands are DNA sequences of a few hundred base pairs that have a high amount (more than 55%) of cytosine and guanine nucleotides. They are associated with about 60% of known human genes, and might be useful in gene identification. The two sequencing consortia^{1,2} again applied different criteria to count CpG islands (Box 1), and these differences probably produce a total that is higher — by an unknown amount — for chromosome 22's long arm. Even so, chromosome 21 appears to be even poorer in CpG islands than in genes when compared with chromosome 22.

The chromosome 22 sequencing consortium suggested that its identification of 545 genes on the long arm was low — a conclusion based in part on the fact that 271 of the 553 identified CpG islands have not yet been associated with genes. In fact, nearly all of the 115 conservatively predicted CpG islands on chromosome 21's long arm are associated with genes. Analysis of both chromosomes using the same methods will help to determine the accuracy of identifying genes by counting CpG islands.

The chromosome 21 sequencing consortium also compared the chromosome 21 sequence with data in the available mouse genome database. No new conserved syntenies — regions where the same genes are 'linked' on chromosomes in different species — were identified. The previously known conserved syntenies are with mouse chromosomes 10, 16 and 17. The chromosome 21 consortium suggests that discrepancies in the gene order predicted by comparing the sequence to mouse

Box 1 Finding genes in a DNA haystack

The two groups^{1,2} involved in sequencing chromosomes 21 and 22 used a similar combination of methods to search the sequences for genes. But they set different criteria to arrive at the numbers of genes and CpG islands — regions of DNA with more than 55% of cytosine and guanine nucleotides, which often mark the 5' ends of genes.

Sixty per cent of mammalian genes are reported to have a CpG island near their 5' end. But the percentage of CpG islands associated with genes is unknown, and can only be determined by knowing the total number of genes. Chromosome 22's long arm² is reported to have 553 CpG islands. The corresponding number is not given for chromosome 21. Rather, the total of 115 CpG islands reported for chromosome 21 includes only the subset that are not associated with repetitive DNA elements. This lowers the CpG island total on

chromosome 21 by an unknown amount.

The chromosome 21 consortium¹ used a conservative criterion for identifying genes amid DNA sequences. This criterion demanded that regions matching the short sequences representing the transcripts of genes should show evidence of having been spliced from multiple protein-coding gene portions (exons).

This would lower the predicted number of genes relative to the calculation for chromosome 22 (ref. 2), which included matches of these expressed sequence tags (ESTs) to single putative exons. Genes with large 3' untranslated regions or large introns (non-coding parts of genes), or those represented in EST databases only by untranslated sequences, are likely to be excluded by the chromosome 21 criterion.

However, the estimate of 225 genes on the long arm of chromosome 21 includes those that are identified only by

computer algorithms that can predict exons from a variety of sequence features (*in silico* prediction). Those with a gene-like structure identified by two or more algorithmic methods were added into the chromosome 21 gene total. Gene-prediction programs can give high false-positive rates of exon prediction, even in combination, and could inflate the gene number.

For chromosome 22, genes predicted only by algorithmic methods do not contribute to the total of 545 genes. But the analysis of chromosome 22 included a projection that was corrected for false positives resulting from *in silico* predictions. If genes predicted only by algorithm were included, the chromosome 22 total might rise by about 100. So the differences in the gene-identification strategies used by the two consortia will tend to cancel each other out. The degree to which this is true could be determined by re-analysis of each sequence using the other method. **R.H.R.**

gene-linkage maps may result from the differing resolution of these maps. In fact, the higher-resolution physical map⁶ of mouse chromosome 10 shows that all 24 genes known to be shared between mouse chromosome 10 and human chromosome 21 occur in the same order. The high degree of conservation between human and mouse is important, because comparing the two sequences — as more of the mouse sequence becomes available — is likely to increase our ability to pick out genes and other significant features from the welter of sequence information.

The availability of the chromosome 21 sequence will have an immediate impact on the study of human single-gene disorders. For example, the genes responsible for five of those monogenic disorders that map to chromosome 21 — including two forms of deafness, Usher and Knobloch's syndromes — have not yet been identified. But having the complete sequence will obviate the labour-intensive step of identifying candidate genes. The genes responsible for these disorders are likely to be found rapidly.

But the greatest impact of the chromosome 21 gene catalogue will be in assessing

the contributions of specific genes to traits seen in Down syndrome. The small number of genes on chromosome 21 is likely to be part of the reason why the presence of three copies of this chromosome — unlike so many chromosome defects — is not fatal at a very young age, or even before birth. Yet there are varying ideas about which genes are associated with particular features of Down syndrome, and the mechanisms by which an imbalance in the number of genes might produce the more than 80 physical and mental disorders that can be seen in this trisomy. Obtaining a comprehensive catalogue of the genes on chromosome 21 has been a goal of Down syndrome researchers for many years, and is realized in this landmark contribution. ■

Roger H. Reeves is in the Department of Physiology, The Johns Hopkins University School of Medicine, 725 North Wolfe Street, Baltimore, Maryland 21205-2105, USA.

e-mail: rreeves@welch.jhu.edu

1. The Chromosome 21 Mapping and Sequencing Consortium *Nature* **405**, 311–319 (2000).
2. Dunham, I. *et al.* *Nature* **402**, 489–495 (1999).
3. Wang, S. Y. *et al.* *Genome Res.* **9**, 1059–1073 (1999).
4. Deloukas, P. *et al.* *Science* **282**, 744–746 (1998).
5. The MHC Sequencing Consortium *Nature* **401**, 921–923 (1999).
6. Wiltshire, T. *et al.* *Genome Res.* **9**, 1214–1222 (1999).

Astronomy

The comet no one saw

Michael F. A'Hearn

With all the work that is going into searching for near-Earth asteroids (NEAs), one might think that any comets approaching Earth would also be discovered. It is expected that, with some increase in the current effort, all near-Earth objects greater than 1 km in diameter will be catalogued within a couple of decades. So how is it possible that a moderately bright comet could be discovered by a spacecraft designed for quite a different purpose and yet never be seen by NEA surveys or, for that matter, by any other observer?

The Solar Wind Anisotropies (SWAN) instrument on the Solar and Heliospheric Observatory (SOHO) spacecraft was designed to study the solar wind that separates our immediate neighbourhood from interstellar gas. But according to Mäkinen *et al.*¹ (writing on page 321 of this issue), images taken by SWAN in the summer of 1997 reveal the presence of five comets in the southern sky, one of which had not been seen by any observers when it was bright.

The SWAN system is designed to observe resonant emission at ultraviolet wavelengths from hydrogen atoms, or Lyman- α emission (Fig. 1), and is therefore also sensitive to comets, which emit copious Lyman- α radiation. SWAN is not easy to use for discovering comets because it has limited resolution, but Mäkinen *et al.* use a neural-network algorithm to find objects in very noisy data. In this way they are able to identify bright comets, such as Hale-Bopp and Hyakutake. The discovery of a new comet purely by its Lyman- α emission should not really be that surprising. The ability of the SOHO satellite to scan the entire sky on a daily basis makes it a unique observing tool.

In 1994, NASA asked Gene Shoemaker to head a committee to devise a realistic programme for detecting asteroids that threaten Earth. They came up with a scheme to detect all such asteroids larger than 1 km in diameter within a decade. It was quickly realized that finding all asteroids above a given size that can cross Earth's orbit is feasible only because the asteroids never move far away from Earth and return close to Earth's orbit every few years². Comets that cross Earth's orbit, on the other hand, have a much wider variety of orbits. Most comets have much larger orbits than asteroids, with correspondingly longer orbital periods. Moreover, the long-period comets rarely travel in the same orbital plane as the planets and asteroids. Finding such objects is much more difficult and NEA search strategies are not optimized for the detection of comets.

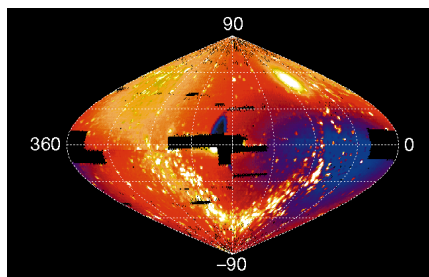


Figure 1 SWAN full-sky map at Lyman- α wavelengths. The equator depicts the plane along which the SOHO spacecraft moves. The black regions along the equator are shadows from the spacecraft. This image was taken in March 1996 when comet Hyakutake was at its closest approach to the Earth, as seen here from the bright white spot in the upper right portion of the map. Mäkinen *et al.*¹ describe similar images taken by SWAN in the summer of 1997 that reveal the presence of an undiscovered comet — it had not been observed by anyone on Earth when it was bright. (From ref. 9.)

Unfortunately, comets probably represent some of the largest hazardous objects we face^{3,4}, but a sensible early warning system has yet to be devised.

Since the beginning of 1997, 22 newly discovered, independent comets have come as close to the Sun, or closer to the Sun, than the comet discovered by Mäkinen *et al.*¹, and 11 have passed inside Earth's orbit⁵. Of these 22 comets, six were discovered by the LASCO coronagraph on the SOHO spacecraft and were quite bright because they were close to the Sun when they were discovered. In addition, nine comets were discovered in the course of systematic surveys using charge-coupled device cameras, while the remaining seven comets were discovered visually by amateur astronomers. All seven comets discovered by amateurs were brighter than the brightest comets identified in professional surveys. But they were bright enough to be detected by these surveys for many months before they became visible to visual observers. It is still the case that amateurs discover many of the new, bright comets, particularly out of the orbital plane of the Solar System. Success is less dependent on what instrument you use than on where you are looking.

The comet discovered by Mäkinen *et al.*¹ was comparable in brightness to some of the fainter comets discovered visually by amateurs, and was brighter than any of the comets discovered by professional surveys. No visual observations were reported but the visual brightness can be estimated from the



100 YEARS AGO

We learn from the *Electrician* that an instrument called the telephonograph, which is a modification of the phonograph, was recently inspected and tested by the German Postmaster-General and several engineers. Its inventor, Herr Paulsen, a Dane, has replaced the wax cylinder of the Edison phonograph by a steel band, and the style by a magnet energised by a telephone. Currents transmitted by the telephone pass through the electromagnet and create consequent poles on the steel band, and more or less the converse operation is employed for reproducing the sound. A long line can, of course, intervene between the transmitting telephone and the phonograph itself, and it is suggested that a telephone subscriber on leaving his office can set such a telephonograph to receive telephoned messages during his absence.

From *Nature* 17 May 1900.

50 YEARS AGO

Prof. F. T. Lewis, of Harvard Medical School, has recently published an interesting comparison of the shapes of soap bubbles and of undifferentiated cells (*Proc. Amer. Acad. Arts and Sci.*, 77, No. 5, 149; 1949). He suggests that the shapes of such cells in bodies of many cells is dependent on the following five major variables.

(1) Geometrical necessity: for example, if cells are of a uniform size they will tend towards 14-hedra, though special features may reduce the facet-number below this figure. The peripheral layer of mass of 14-hedra will consist of 11-hedra, each of the 11-hedra making one surface of contact with the surrounding medium instead of four with surrounding polyhedra. (2) Surface tension and other forces tending to produce minimal surface area: with bubbles, surface tension can produce examples of unstackable combinations of twelve, fourteen, fifteen and sixteen facets which, if isolated and regular, have the least surface for volume. None of these shapes has yet been found among cells. (3) Cell division: this must involve forces opposed to surface tension, since division increases the total cell surface area. (4) Turgor: this again is opposed to surface tension. (5) Organisation: under this general heading Prof. Lewis includes all the processes whereby growth and division are "controlled to produce filaments, fibres, membranes or compact masses as may be proper for the organism".

From *Nature* 20 May 1950.

brightness of the ultraviolet Lyman- α emission (based on data from comets that have been observed at both wavelengths). Orbital calculations^{1,6} suggest that this comet was brighter than the brightest comet discovered by the professional surveys for at least six months. Why was it not seen at visible wavelengths? Does this imply that there are many other comparable comets just waiting to be discovered? And, if so, should we care?

We care about the efficiency of discovering comets for many reasons. The most obvious scientific reason is that we cannot find out where comets come from and how big they are unless we can reduce the biases in our observations. These biases are due to selective discovery in certain parts of the sky and to varying orbital and physical characteristics of the comets. For our peace of mind, it is important to know whether comets represent 10% of the potential large impacts on Earth (as is commonly thought) or a much larger fraction. Many people argue that comets dominate at the largest sizes and that the observations set only a lower limit to this fraction.

It has long been recognized that selection effects influence the discovery of comets, but the magnitude of these effects is not well understood. A few years ago, Brandt *et al.*^{7,8} proposed a systematic search for small (100-metre) comets using a technique similar to SWAN. But instead of studying emission from hydrogen they suggested using emission from the OH radical (the other fragment produced when cometary water molecules are broken apart by sunlight) as a characteristic marker of cometary activity. Their proposal suffered from inadequate satellite time, and no comets were found.

The comet found by Mäkinen *et al.* is a prototypical example of something missed by other people because of the strong selection effects in those other searches. A systematic search for comets in the inner Solar System, using the known characteristics of comets and searching the entire sky with high frequency, would be extremely useful. Although observations of emission by the OH radical would have less of a problem with background noise, a systematic search by SWAN using longer integration times would almost certainly turn up additional comets. It would add tremendously to our knowledge of the physical and orbital properties of comets and would provide an insight into the real statistical frequency of cometary impacts on Earth.

Michael F. A'Hearn is in the Department of Astronomy, University of Maryland, College Park, Maryland 20742-2421, USA.
e-mail: ma@astro.umd.edu

1. Mäkinen, J. T. T. *et al.* *Nature* **405**, 321–322 (2000).
2. Gehrels, T. (ed.) *Hazards due to Comets and Asteroids* (Univ. Arizona Press, Tucson, 1994).
3. Grieve, R. A. F. & Shoemaker, E. M. in *Hazards due to Comets*

and *Asteroids* (ed. Gehrels, T.) 417–462 (Univ. Arizona Press, Tucson, 1994).

4. Bailey, M. E. *et al.* in *Hazards due to Comets and Asteroids* (ed. Gehrels, T.) 479–533 (Univ. Arizona Press, Tucson, 1994).
5. Marsden, B. G. & Williams, G. V. *Catalogue of Cometary Orbits 1999 13th edn* (Minor Planet Center, Cambridge, MA, 1999).
6. Marsden, B. G. *Minor Planet Electronic Circular 1999-X07*

(Minor Planet Center, Cambridge, MA, 1999).

7. Brandt, J. C. *et al.* *Earth, Moon and Planets* **72**, 243–249 (1996).
8. Brandt, J. C. *et al.* in *Completing the Inventory of the Solar System* (eds Rettig, T. W. & Hahn, J. M.) 289–297 (Astron. Soc. Pacific, San Francisco, 1996).
9. <http://www.geo.fmi.fi/PLASMA/SWAN>

Cancer

Checkpoint for invasion

Lance A. Liotta and Timothy Clair

Both benign and malignant tumours grow in an uncontrolled way. But it is only cells of malignant tumours that invade surrounding tissues and travel to distant organs (metastasize). Conventional wisdom used to hold that invasion and metastasis are late events — often 'too late' — in the clinical course of a patient's cancer. However, we now know that invasion can be both early and clinically 'silent'. An understanding of the molecular basis for this aggressiveness could lead to therapies that block the transition of a tumour from benign

to malignant, and keep local disease in check. Taguchi and colleagues¹, writing on page 354 of this issue, have now identified proteins called RAGE and amphoterin as a receptor–ligand pair in a molecular checkpoint that regulates not only the invasiveness but also the growth and movement of tumour cells — the trio of characteristics required for malignancy.

The threat of tumour invasiveness is exemplified by the fact that brain cancer does not need to metastasize to kill a patient. The growth of a brain tumour mass in the

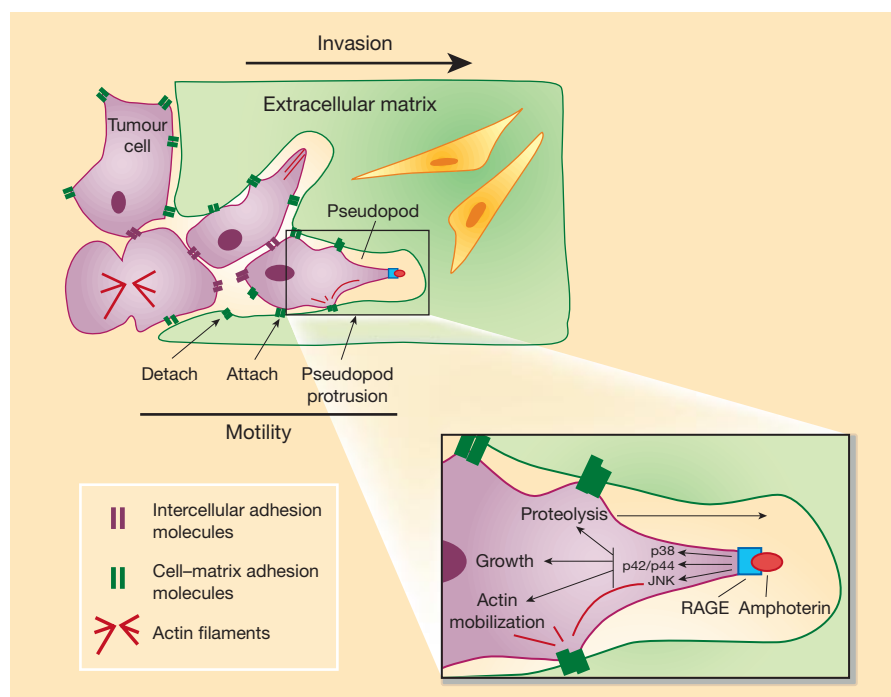


Figure 1 Spatial and temporal regulation of cellular invasion of the extracellular matrix. Invasion can be normal (for example, in the case of the outgrowth of neuronal protrusions during brain development) or malignant (for example, in the case of tumours). Top, invasion can be viewed as cellular motility coupled to regulated adhesion and detachment (from the extracellular matrix) and proteolysis (of extracellular matrix molecules). The advance of pseudopods (extensions) of the cell — driven by the formation of actin polymers, a component of the cytoskeleton — may require the action of cell-surface protein-degrading enzymes, as well as other enzymes, receptors and activators. Extracellular matrix degradation must be balanced by antiproteolysis (that is, processes that inhibit proteolytic enzymes) to allow for adhesive traction. Bottom, signal-transduction pathways allow the individual cell to move between phases of pseudopod protrusion, extracellular matrix degradation, antiproteolysis, adhesion and detachment. These pathways split at the level of the mitogen-activating protein kinases JNK, p38 and p42/p44. Blocking the interaction between amphoterin and RAGE suppresses these pathways, as shown by Taguchi *et al.*¹.

confined area of the skull causes compression damage; in addition, local invasion by brain tumour cells can destroy surrounding, normal brain tissue. In many cases, brain tumour cells can move away from the primary tumour to reach other sites within the brain. Such insidious invasive behaviour may represent the inappropriate use of a programme responsible for the outgrowth of neuronal protrusions called neurites during normal neuronal development. Indeed, cancer invasion in general may be a deregulated form of a physiological invasion process required for neuronal wiring in the embryo, tissue remodelling, the formation of blood vessels, and healing².

Amphoterin is a key protein in normal neurite outgrowth. It is a heparin-binding protein that is abundant in extracellular regions of the developing brain and other organs. Antibodies that recognize amphoterin block neurite outgrowth under experimental conditions; so, interactions of amphoterin with neuronal surfaces appear to be required for the extension of neuronal processes. Amphoterin's receptor on the cell surface is a protein called RAGE (for 'receptor for advanced glycation end products')³. RAGE is a receptor for many different ligands, and is a member of the immunoglobulin superfamily of cell-surface molecules. It gains its name, and was first identified, because it recognizes potentially damaged, glycated proteins (that is, those with carbohydrate polymers attached to them) that accumulate during diabetes. Amphoterin and RAGE localize together at the leading edge of advancing neurites during embryonic development³. Taguchi *et al.*¹ recognized the implications of this result for pathological processes such as cancer invasion.

The complete set of characteristics of a malignant tumour is induced by a variety of cellular programmes and pathways, many of which are not yet fully defined. Faced with this complexity, the best way to link a molecule causally to malignancy is to start with a cell that is already malignant, and to attempt to block the molecule or pathway of interest. This was the tack taken by Taguchi *et al.*¹, who used several approaches to block the RAGE–amphoterin axis in C6 glioma brain tumour cells. Inhibitory strategies included administering the soluble form of the ligand-binding domain of RAGE, of anti-RAGE antibodies, or of anti-amphoterin antibodies, or introducing defective forms of RAGE into C6 glioma cells. *In vitro* and in animal models of cancer, all of these treatments significantly inhibited the growth, motility and local invasion of tumour cells, as well as metastasis of the cells to the lungs. The treatments even inhibited the spontaneous growth of papillomas (benign skin cancers) in mice overexpressing the v-Ha-ras oncogene. How might the

RAGE–amphoterin pathway have these effects?

Let's take a look first at invasion. Regulation of the molecular events necessary for invasion — whether physiological or malignant — involves spatial and temporal coordination, as well as cyclic on–off processes, at the level of individual cells (Fig. 1). Motility, coupled with regulated, intermittent adhesion to the extracellular matrix and degradation of matrix molecules, allows an invading cell to move through the three-dimensional matrix. At the leading edge of the motile cell, receptor–ligand and proteolysis–antiproteolysis complexes coordinate sensing, protrusion, burrowing and traction of the cell^{4,5}.

It was already known that amphoterin at the cell surface can act as a nucleating site for generation of the protein-degrading complex plasmin⁶. This complex can activate matrix metalloproteinases (MMPs), which are enzymes that degrade extracellular matrix molecules⁵. Taguchi *et al.*¹ now report that blocking RAGE results in decreased activity of MMP-2 and MMP-9 — molecules previously associated with invasion of both cancer cells and neurites. Localized proteolysis of matrix molecules may loosen up, or open up, the dense meshwork of matrix molecules being invaded. Proteolysis at the migration front may also liberate previously bound growth factors or motility-stimulating molecules.

The RAGE–amphoterin complex also suppresses tumour growth, but how does it coordinate these three inhibitory effects — on growth, on motility and on invasion? Proteins called cytokines, as well as proteins found in the extracellular matrix, trigger signalling cascades that regulate both cell migration and proliferation. Bifurcation of this signalling pathway occurs at the level of mitogen-activated protein kinase (MAPK) signalling modules. Three coexisting mod-

ules — p38^{MAPK}, JNK and p42/p44^{MAPK} — exchange signals between the cell surface, the cytoskeleton and the nucleus.

When a ligand stimulates the cell through that ligand's receptor, some or all of these MAPK modules can be activated, directly or indirectly, as can the small GTP-hydrolysing proteins (GTPases) Ras, Cdc42, Rac and Rho^{7,8}. Activated MAPK modules propagate signals downstream into the nucleus to activate genes encoding growth inducers, MMPs and adhesion receptors. In parallel, these modules elicit further events that modify the myosin and actin filaments of the cytoskeleton. So all three MAPK modules can act as relay stations for the regulation of growth, motility and invasion. Taguchi *et al.*¹ show that RAGE–amphoterin acts simultaneously through all three MAPK modules, explaining how blocking RAGE will experimentally suppress all components of the malignant phenotype.

'Signal-transduction therapy' is a treatment strategy in which key, hyperactive cellular signalling pathways that cause disease are targeted. The trick is to find a rheostat in the cell's circuitry that is not bypassed by collateral or compensatory paths. The RAGE–amphoterin pathway may well fulfil these criteria.

Lance A. Liotta and Timothy Clair are in the Laboratory of Pathology, National Cancer Institute, National Institutes of Health, Building 10/Room 2A33, Bethesda, Maryland 20892, USA.
e-mails: lance@helix.nih.gov
timclair@helix.nih.gov

1. Taguchi, A. *et al.* *Nature* **405**, 354–360 (2000).
2. Liotta, L. A., Steeg, P. S. & Stetler-Stevenson, W. G. *Cell* **64**, 327–336 (1991).
3. Hori, O. *et al.* *J. Biol. Chem.* **27**, 25752–25761 (1995).
4. Lauffenburger, D. & Horowitz, A. *Cell* **84**, 359–369 (1996).
5. Murphy, G. & Gavrilovic, J. *Curr. Opin. Cell Biol.* **11**, 614–621 (1999).
6. Parkkinen, J. *et al.* *J. Biol. Chem.* **268**, 19726–19738 (1993).
7. Denhardt, D. *Biochem. J.* **318**, 729–747 (1996).
8. Banyard, J. *et al.* *Oncogene* **19**, 580–591 (2000).

Palaeoceanography

Nutrients in the glacial balance

Peggy Delaney

What are the natural controls on the concentration of carbon dioxide in Earth's atmosphere? The question is a compelling one given the rising levels of CO₂ from human activities and the prospect of global warming. The past has many lessons for us, and the latest comes from Elderfield and Rickaby on page 305 of this issue¹. Their analysis hinges on inferences about oceanic nutrient levels during the most recent period of intense glaciation, the Last Glacial Maximum (LGM), around 22,000 years ago. Nutrient levels can tell us about the level of phytoplankton primary

productivity, and so CO₂ uptake, in the ocean; this in turn can help account for the amount of CO₂ in the atmosphere.

Climate modelling is one approach to tackling the 'CO₂ issue'. But it has its limitations, so attention has also turned to times in the geological past when concentrations of CO₂ and climate were very different. Because of the immense reservoir of inorganic carbon dissolved in the ocean (equivalent to over 60 times the amount of atmospheric carbon), explanations for large and geologically rapid changes in atmospheric CO₂ centre on the oceanic carbon cycle.

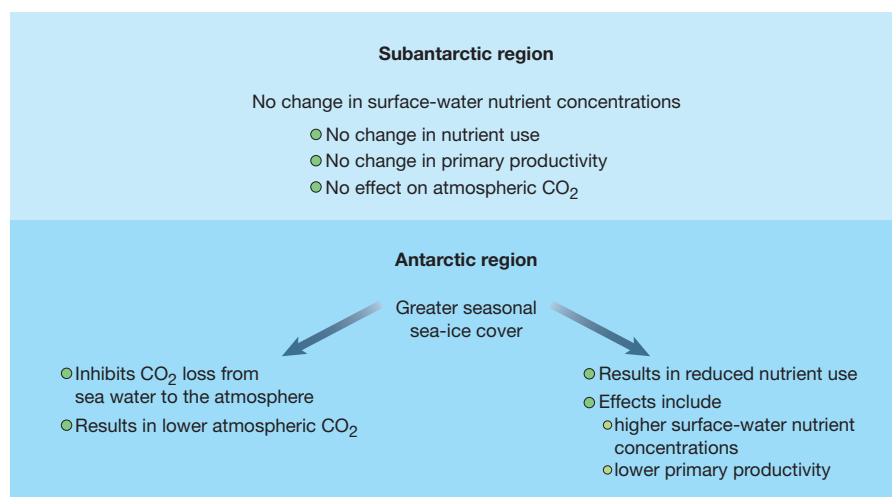


Figure 1 Primary productivity in the Southern Ocean during the Last Glacial Maximum (LGM), as surmised by Elderfield and Rickaby¹ and relative to today. The authors interpret core records of the cadmium/calcium ratio in foraminiferal calcite in terms of surface-water phosphate, and therefore nutrient, history. They conclude that the situation in the subantarctic region of the Southern Ocean during the LGM was much the same as it is today, but that in the Antarctic region productivity was lower than now. This conclusion can be reconciled with the lower levels of atmospheric CO₂ during the LGM by taking into account the effects of greater sea-ice coverage. The sea ice would have greatly inhibited CO₂ loss to the atmosphere.

Ice-core records of trapped gases indicate that levels of CO₂ in the atmosphere were much lower (by about 90 parts per million by volume, or about one-third) during the LGM. Are changes in nutrient use in the vast Southern Ocean the key to understanding lower atmospheric CO₂ levels during the LGM? Elderfield and Rickaby offer a novel interpretation of primary-productivity history in the Antarctic and subantarctic regions of the Southern Ocean (Fig. 1). They suggest that Antarctic productivity was lower then than it is now, and that subantarctic productivity was unchanged. A surprising twist is that if the effects of sea-ice cover are taken into account, lower Antarctic productivity can be reconciled with lower levels of atmospheric CO₂.

In the sunlit surface waters of the oceans, microscopic organisms transform inorganic dissolved macronutrients such as phosphorus, nitrogen and carbon into organic matter. Some of that organic matter is then transported in particles to deeper waters. Some primary producers build hard parts of silicon, so this element also behaves like a nutrient. Of particular interest are regions in the modern ocean where macronutrient concentrations are not reduced to the extremely low levels characteristic of most of the surface ocean — these are the 'high nutrient, lower (than expected) chlorophyll' regions, such as the Southern Ocean.

Changes in nutrient use — the fraction of available nutrients used in primary production — in these areas have the largest potential to radically alter carbon exchange with the atmosphere. Explanations for the lower levels of CO₂ during the LGM often invoke higher productivity in part or all of

the Southern Ocean. Surprisingly, Elderfield and Rickaby turn this view on its head. From evidence for higher levels of nutrients in surface waters, and so lower nutrient use, they infer that primary productivity in the Antarctic was less than it is now.

The conclusions reached by Elderfield and Rickaby stem from new insights into the relationship of cadmium and phosphorus in sea water, and into influences on the cadmium/calcium ratio in planktonic foraminifera — these are microscopic zooplankton, which produce hard parts of calcite that become preserved in marine sediments. Ed Boyle paved the way in this area, developing the use of Cd/Ca ratios in foraminiferal calcite as proxy tracers of seawater nutrient concentrations (see refs 2 and 3 for example). Deciphering oceanic chemical history depends on proxies such as this, where the composition of material from sediment cores provides records of the past chemical composition of the ocean. The Cd/Ca ratio in foraminiferal calcite is proportional to that in the sea water in which the organism grew. Calcium is effectively constant with salinity throughout the ocean, so the Cd/Ca ratio in sea water depends on Cd. Seawater Cd, in turn, varies with the nutrient element P.

Elderfield and Rickaby¹ conclude that the Cd/P relationship can be best explained by preferential uptake of Cd relative to P (by a factor of 2.5) as a result of biological activity in surface waters. The degree of nutrient use — how much dissolved P is taken up in biological activity — determines seawater Cd/P, as Cd is more rapidly depleted than P. With their new interpretation of Cd/P systematics in sea water, Elderfield and Rickaby argue

that they can more accurately translate Cd/Ca ratios measured in foraminiferal calcite to P concentrations in past surface waters.

Temperature is also an important consideration. In an earlier paper⁴, Rickaby and Elderfield concluded that the Cd/Ca ratio in foraminiferal calcite is temperature-sensitive, with higher temperatures resulting in a higher Cd/Ca ratio for a given concentration of Cd in sea water. Records of planktonic foraminiferal Cd/Ca from subantarctic cores (for instance that from core E48-22; see Fig. 4 of the paper¹ on page 307) show large reductions in the ratio during glacial intervals. Records from more southerly Antarctic cores (such as MD 80-304; Fig. 4 again), by contrast, show little glacial–interglacial change in Cd/Ca. When the effects of the lower temperatures of the LGM are taken into account, these Cd/Ca records indicate little glacial–interglacial change in surface-water nutrient concentrations in the subantarctic. But they imply that levels of P were higher in Antarctic surface waters during the LGM, indicating a lower degree of nutrient uptake and therefore lower productivity (Fig. 1).

Elderfield and Rickaby¹ also take on the challenge of reconciling their new interpretation of the Cd/Ca ratio with other proxies for nutrient use in the Southern Ocean. They make a persuasive case that most other records (nitrogen, carbon and silicon) are consistent with their interpretation. The one apparently contradictory tracer, nitrogen isotopes from cores in the Antarctic region, may instead reflect the anomalous effects of sea-ice cover on these isotopes. Finally, the authors conclude that increased sea-ice cover, which reduces biological productivity, acts as a barrier to air–sea exchange and accounts for the lower atmospheric CO₂ during the LGM (Fig. 1).

The strengths of this work¹ are that it provides a more systematic, global explanation for the Cd/P relationship in sea water, and that it integrates interpretations of various nutrient proxy records for the Southern Ocean. But the field awaits two further developments: an integrated, multi-proxy tracer approach on a core or cores from this critical oceanic region, and longer records that span several glacial–interglacial cycles.

Finally, what are the broader implications of Elderfield and Rickaby's insights into the Cd/P relationship and the effects of temperature on the foraminiferal Cd/Ca ratio? As they themselves point out, the new interpretation of the Cd/P relationship will not significantly affect the use of the Cd/Ca ratio in calcite as a tracer of deep-water circulation patterns. Nor will the influences of temperature on the ratio affect studies of deep-water properties for the recent geological past, when deep-water temperatures were similar to those today. But they must be taken into

account for time intervals when oceanic deep waters were much warmer than now, as must changes in oceanic Cd and P budgets. ■

Peggy Delaney is in the Ocean Sciences Department and the Institute of Marine Sciences, University of California, Santa Cruz, California 95064, USA.

e-mail: delaney@cats.ucsc.edu

1. Elderfield, H. & Rickaby, R. E. M. *Nature* **405**, 305–310 (2000).
2. Boyle, E. A. & Keigwin, L. D. *Earth Planet. Sci. Lett.* **76**, 135–150 (1985/86).
3. Boyle, E. A. *Paleoceanography* **3**, 471–489 (1988).
4. Rickaby, R. E. M. & Elderfield, H. *Paleoceanography* **14**, 293–303 (1999).

Structural biology

Bioluminescence illuminated

Franklyn G. Prendergast

On page 372 of this issue¹, Head and colleagues report the tertiary structure of the fascinating photoprotein aequorin. The protein is activated by calcium. It is of interest both because it is widely used in the laboratory for tracking cellular calcium (and more), and for its natural function — bioluminescence in the jellyfish *Aequorea aequorea*.

Bioluminescence is the phenomenon in which visible light is emitted by an organism. The biochemical reaction involves oxidation of a substrate (termed a luciferin) by an enzyme (a luciferase), molecular oxygen usually being the oxidant. Many terrestrial and marine organisms bioluminesce, and there are different classes of luciferases and luciferins. Aequorin is a member of a unique

class of bioluminescent proteins, photoproteins², which resemble enzymes trapped in an intermediate state. Photoproteins require only the further presence of Ca^{2+} ions to trigger completion of the intramolecular chemical reaction and emission of a photon. The results of Head and colleagues¹ provide a splendid insight into aequorin, and allow plausible speculation on the mechanisms underlying the Ca^{2+} -triggered bioluminescence.

Originally aequorin was not classed as an enzyme. Shimomura³ was the first to show that by removing Ca^{2+} from the protein, and incubating the resulting apo-aequorin with synthetic luciferin (an imidazopyrazinone) in the presence of molecular oxygen and a reducing agent, the protein could be

recharged repeatedly and reversibly — that is, it could be made to function as an enzyme. The turnover rate *in vitro* is very low. Nothing is known of how the process works *in vivo*. But the jellyfish itself definitely regulates its bioluminescence by nerve-controlled Ca^{2+} release and sequestration in aequorin-containing cells; so presumably it can control aequorin's enzymatic activity as well.

Several other members of the coelenterate group of marine invertebrates have Ca^{2+} -activated photoproteins. Among them are *Mnemiopsis* (photoprotein called mnemiopsin), *Obelia* (obelin), *Beröe* (berovin), *Halistaura* (mitrocomin) and *Phialidium* (phialidin). (Ward and colleagues provided the earliest characterization of two of these other photoproteins⁴.) They all use the same imidazopyrazinone as a substrate, which is why this luciferin is called coelenterazine. Sequence similarity among the proteins is such that their tertiary structures can all be inferred. It should then be possible to evaluate Head and colleagues' proposals as to the structural basis for stabilization of the hydroperoxide, and maybe the hydroperoxy anion, of coelenterazine in photoproteins (Fig. 1). Their proposed role for the carboxy-terminal proline residue is particularly provocative.

The precise catalytic mechanism in the photoprotein remains very much in question. The residues constituting the (putative) active site are not evident from the tertiary

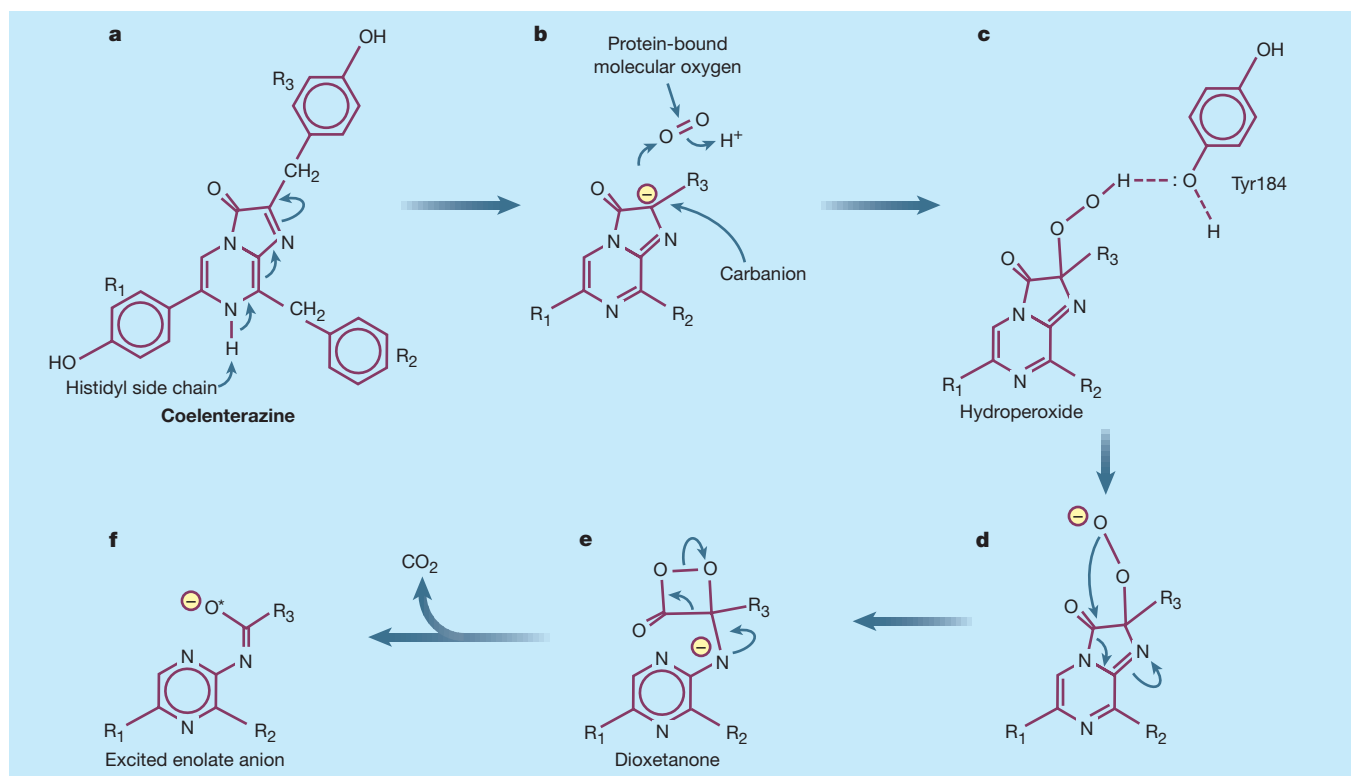


Figure 1 Possible chemical reactions underlying bioluminescence of aequorin. a, Distortion of the ground state of coelenterazine by the aequorin protein leads to the formation of a reactive carbanion. b, The carbanion attacks protein-bound molecular oxygen, leading to the hydroperoxide (c) which is stabilized primarily by hydrogen bonding to a tyrosine residue at position 184. d, Binding of Ca^{2+} to aequorin induces a conformational change which allows attack by the hydroperoxide anion on the reactive carbonyl, leading to transient formation of an unstable dioxetanone (e). The dioxetanone readily undergoes scission to yield CO_2 and an enolate anion (f) in an excited state, which returns to the ground state by emission of a photon of blue light.

structure. Indeed, it might be that the protein acts primarily as a protective host for the two co-substrates, coelenterazine and molecular oxygen, which are held in appropriate orientation relative to each other and are stabilized by the protein matrix. Oxidation to the hydroperoxy anion could thereafter occur quite spontaneously because of relatively facile intramolecular generation of a reactive carbanion from the coelenterazine (Fig. 1b).

Formation of the carbanion is the central process in activation, and is probably facilitated by a side chain of a basic amino acid in the protein (for example by a histidyl side chain, as shown in Fig. 1a). Head *et al.* discuss potential roles for (apparently) key histidine residues in aequorin. But in this paper they do not touch on the question of what function, if any, cysteinyl (thiol) moieties play; reversible chemical modification of thiols in *Aequorea* leads to reversible loss of Ca^{2+} -sensitive bioluminescence.

Coelenterazine is also the substrate for non-photoprotein luciferases found in diverse marine species such as *Renilla*, *Ptilosarcus* and *Gaussia*. These enzymes show no sequence similarity to the photoproteins and are not activated by Ca^{2+} ; and



Figure 2 *Aequorea aequorea* — a shining light of biology.

although they are relatively slow compared with most enzymes, their turnover rates are much higher than that of aequorin. Yet these non-photoprotein luciferases also require only oxygen and coelenterazine for bioluminescence, implying that the catalytic mechanism is similar to that in aequorin. Presumably there are also similar substrate-binding motifs — equivalent, maybe, to the common nucleotide-binding motifs found in proteins of dissimilar sequence.

Despite the availability of this tertiary structure, the mechanistic basis for Ca^{2+} triggering of bioluminescence in aequorin is not yet solved. Head *et al.* speculate reasonably that relatively small and local Ca^{2+} -induced conformational changes would disrupt the tyrosine (phenol)–hydroperoxide hydrogen bond (Fig. 1c), thereby facilitating nucleophilic attack by the hydroperoxy anion on the reactive carbonyl of the substrate (Fig. 1d). Thereafter the intramolecular reaction sequence depicted in Fig. 1 would proceed readily. Given the energetics of hydrogen bonding, thermally driven structural fluctuations should suffice to trigger Ca^{2+} -independent light emission in aequorin. They indeed do so, although the rate of such emission in *Aequorea* is less than a millionth of light emission triggered by Ca^{2+} binding².

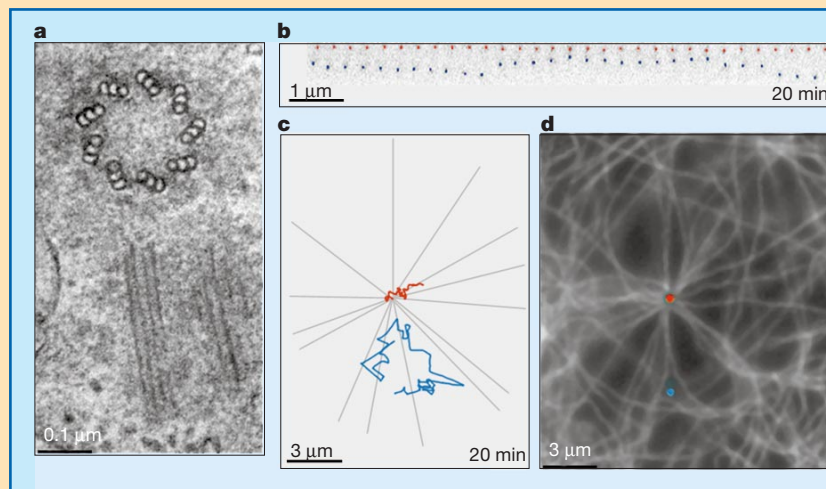
Interestingly, other divalent and trivalent cations can evoke aequorin bioluminescence, some less so and some more so than Ca^{2+} . The Ca^{2+} -binding sites known as EF-hands, such as those in *Aequorea*⁵, bind a variety of multivalent cations which can sometimes act as isomorphous replacements for Ca^{2+} . Such promiscuity supports

Cell biology

Centrioles go for a stroll

Centrioles are arguably the most enigmatic cell organelles. In most animal cells, a pair of centrioles resides in the centrosome — a macromolecular complex that organizes the microtubule system (part of the cell's internal skeleton). The two centrioles, a mother and a daughter, “are positioned at right angles to each other”. This quote from a popular textbook epitomizes the idea of a fixed arrangement of the two bodies within the centrosome (see part a in the figure: at the top is an end-on view of one of the centrioles, while at the bottom is a longitudinal view of the other centriole). Deviations from a close apposition are considered exceptional.

Not so, say Matthieu Piel and colleagues (*J. Cell Biol.* **149**, 317–330; 2000). They have been following the behaviour of centrioles throughout the cell cycle (the time between successive cell divisions) in living cells. Centrioles are difficult to see by standard light microscopy, so Piel *et al.* made use of the fluorescent tag green fluorescent protein (GFP). They fused GFP to centrin, a small calcium-binding protein present in the lumen of the centriole. Fluorescence — a marker of the amount of GFP-tagged centrin — is always more intense in the mother centriole. So, the intensity of fluorescence seems to be indicative of the maturity of a centriole. This allowed the authors to show that a daughter centriole remains ‘immature’ until it duplicates and itself becomes a mother. The observation also indicates that one centriole may not be able to replace the other functionally, but this needs to be confirmed.



The most striking observation, however, is that the two centrioles exhibit dramatically different behaviour. For several hours after cell division, the daughter centriole roams extensively through the body of the cell, separating from the mother by many micrometres. Motility gradually subsides at the time of centriole duplication, the start of which coincides with the beginning of DNA replication in the nucleus. But these movements also take place — and are more pronounced — in the absence of nuclei. An example of these movements over a period of 20 minutes is shown in b and c of the figure. While the daughter (blue) runs free, the mother centriole (red) resides in the centre of the microtubule array (d in the figure).

Splitting of the centriole pair has previously

been observed under certain experimental conditions or in specialized cell types, but what the work of Piel *et al.* implies is that centriolar excursions are a normal event during a considerable portion of the cell cycle. Why one centriole should roam the cell after completion of cell division is unknown, as is the mechanism underpinning these movements. But the observations shake the myth of a stereotypical, right-angled orientation of centrioles, and suggest that these curious organelles might have some hitherto unsuspected functions associated with their wanderings.

Manfred Schliwa

Manfred Schliwa is at the Institute for Cell Biology, University of Munich, Schillerstrasse 42, 80336 Munich, Germany.
e-mail: schliwa@bio.med.uni-muenchen.de

the notion of a sort of structurally determined 'hair-trigger', which is activated by metal-ion binding to the Ca^{2+} -binding sites, a hypothesis in keeping with Head and colleagues' structure. However, given this hypothesis, it is peculiar that Mg^{2+} stabilizes aequorin, reducing its Ca^{2+} -independent light emission and slowing the rate of Ca^{2+} -triggered bioluminescence. Why don't Mg^{2+} ions also evoke bioluminescence, differences between Mg^{2+} and Ca^{2+} in coordination, geometry and ionic size notwithstanding?

The Ca^{2+} sensitivity and bioluminescence of aequorin have long made it a wonderful sensor for the detection and quantification of intracellular calcium-ion concentration. In the past few years, the ready availability of synthetic coelenterazine, and recombinant-DNA and gene-transfer technologies, has resulted in several organelle-targeted aequorins that have proved particularly helpful for studying the regulation of intracellular calcium^{6,7}.

In *Aequorea* (Fig. 2), aequorin coexists with the well-known green fluorescent protein (GFP). Arguably, GFP is the single most popular tool currently used in research into both gene expression and a host of

intracellular processes. Now that the tertiary structures of both proteins are known, only imagination limits the possibilities of further developments. Moreover, aequorin can transfer its bioluminescence energy non-radiatively to GFP, despite the lack of a demonstrable physical interaction between the two wild-type proteins *in vitro*. One wonders what sorts of reciprocal structural modifications might be made to afford aequorin-GFP heterodimers? Head and colleagues' paper presages all types of possibilities for this luminescent dynamic duo. ■

Franklyn G. Prendergast is in the Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, 200 First Street SW, Rochester, Minnesota 55905, USA.
e-mail: prendergast@mayo.edu

1. Head, J. F., Inouye, S., Teranishi, K. & Shimomura, O. *Nature* **405**, 372–376 (2000).
2. Blinks, J. R., Prendergast, F. G. & Allen, D. G. *Pharmacol. Rev.* **28**, 1–93 (1976).
3. Shimomura, O., Musicki, B. & Kishi, Y. *Biochem. J.* **251**, 405–410 (1988).
4. Ward, W. W. & Seliger, H. H. *Biochemistry* **13**, 1491–1499 (1974).
5. Kretsinger, R. H. & Nelson, D. J. *Coord. Chem. Rev.* **18**, 29–124 (1977).
6. Rutter, G. A. *et al. Proc. Natl Acad. Sci. USA* **93**, 5489–5494 (1996).
7. Robert, V., De Giorgi, F., Massimino, M. L., Cantini, M. & Pozzan, T. *J. Biol. Chem.* **273**, 30372–30378 (1998).

Materials science

Plastic sandwiches à la carte

Michael D. Ward

As materials science and engineering evolve, we are witnessing new approaches that provide increasingly precise control, at the atomic and molecular levels, over the structure, properties and function of materials. Although this is largely driven by a recognition that familiar bulk properties — mechanical, magnetic, electronic and optoelectronic — depend upon atomic and molecular building blocks, these same properties often differ from the bulk when examined at the nanometre length scale. On page 328 of this issue, Matsumoto *et al.*¹ describe how to build a bulk material that fully exploits the molecular properties of the individual building blocks. Their layered polymer crystals could lead to the design of functional solids in areas such as molecular recognition, separation and catalysis.

One aim of materials engineering is the controlled assembly of carefully designed molecules, or ensembles of molecules, into nanometre-scale structures. In this context it is often tempting to mimic naturally occurring systems that provide structural scaffolds. In this respect, there has also been substantial interest in materials that mimic natural clays, which have a characteristic

layered or sheet-like structure. Matsumoto *et al.*¹ create a synthetic clay by building organized, crystalline, molecular scaffolds, which contain reactive monomers. When exposed to light the monomers polymerize to make robust polymer sheets that are only a single molecule thick. Solid-state polymerization in crystals has been demonstrated before, but this new work shows that the polymerized sheets can be separated and then 'reglued' with a 'molecular adhesive', yielding layered materials in which the spacing between the sheets is controlled by the size of the adhesive.

The work of Matsumoto *et al.* hinges on 'crystal engineering' — the design of crystal structures from molecular principles. Almost 40 years ago, Gerhard Schmidt and co-workers demonstrated that chemical reactions in the solid state depend on the arrangement of molecules in the solid crystal lattice, the most reactive crystals being those for which a minimum amount of atomic or molecular motion is required for them to react². The products of these so-called topochemical reactions, which range from cyclic hydrocarbons with controlled stereochemistry to new conducting polymers, have structures that reflect the crystal symmetry

of the reactant. Furthermore, the constraint imposed by the solid-state lattice typically affords improved reaction selectivity compared with the same reactions performed in solution. The work of Schmidt and colleagues launched a new era of crystal engineering^{3,4}.

Matsumoto *et al.* have put these ideas into practice. They create a crystal lattice from muconic acid, which is a reactive monomer with two negatively charged carboxylate groups at each end. These negative charges are compensated by two alkylammonium cations. This structure naturally generates a crystal with layers of reactive monomers (M) sandwiched between layers of the long alkane chains (A) attached to the alkylammonium ions. These AMA sandwiches stack on top of each other in the repeating sequence ...AMAAMA... (Fig. 1a, overleaf), closely resembling the well-established structure of layered clays, which are composed of charged metal-oxide sheets separated from each other by water molecules and counterions. The metal-oxide clay sheets are structurally robust because of covalent bonding within the inorganic sheet. But the layered structure of the unpolymerized AMAAMA crystals is controlled primarily by non-covalent attractive interactions between the long alkane chains and the tendency for polar and non-polar components to segregate.

When exposed to light, the muconate monomers within each layer polymerize to generate a structurally robust plastic sheet that can be thought of as a more flexible version of the metal-oxide clay sheets (Fig. 1b). Many examples of topochemical polymerization between reactive monomers⁵, including those sandwiched between fairly well-defined inorganic and organic sheets^{6–10}, have appeared over the past 20 years. But Matsumoto *et al.*¹ have taken this one step further by showing that the polymerized sheets can be separated from each other and then reassembled into new polymer sandwiches with different fillings.

After polymerization the polymer sheets remain coated on both sides with alkylammonium ions, but the ions can be removed with acid to generate uncharged polymer sheets (Fig. 1c). Remarkably, these free sheets can be joined together again using any one of a number of alkylammonium ions, which serve as a molecular glue. Using X-ray diffraction, Matsumoto *et al.* show that the distances between the layers of the reassembled plastic clays correspond to the size of the alkylammonium ions they choose as the glue.

In most respects the processes studied by Matsumoto *et al.*¹ mimic the intercalation, separation and reassembly behaviour seen in organic clays¹¹, metal phosphates¹², graphite oxide¹³ and perovskites¹⁴. Questions remain over the detailed structure of the polymer sheet, particularly the role

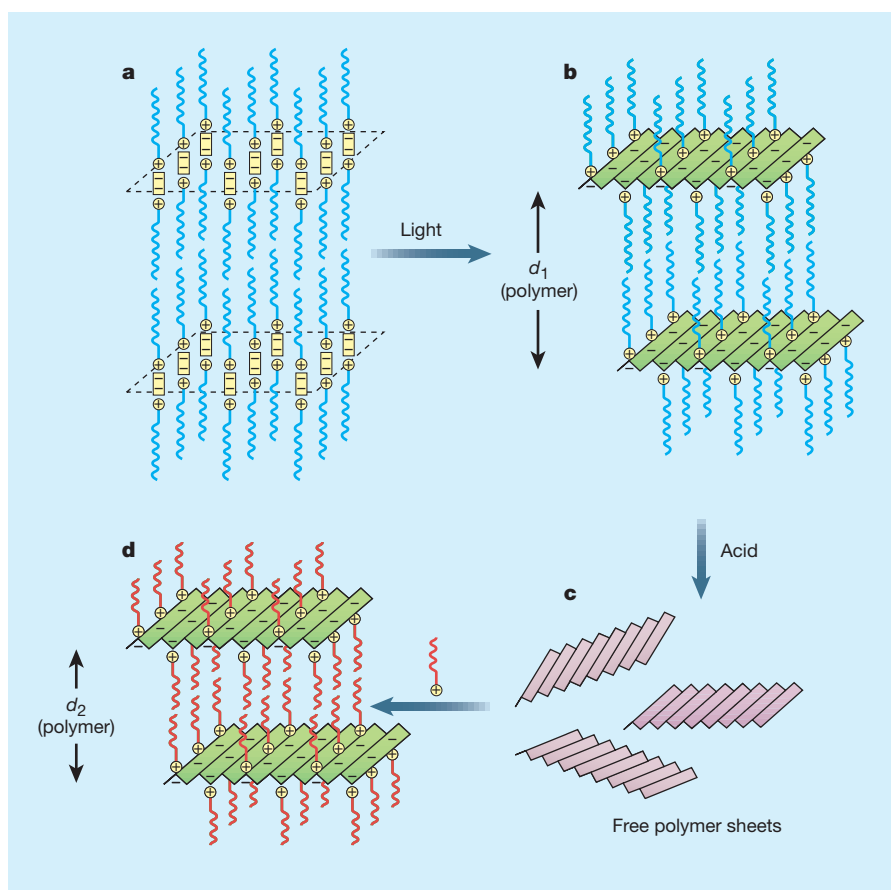


Figure 1 How to build molecular sandwiches with different fillings. **a**, According to Matsumoto *et al.*¹, layers of muconate monomers (yellow) and associated alkylammonium cations (blue chains) will assemble into a crystalline material with layer-to-layer spacings (d_1) governed by the length of the alkane chains of the cations. **b**, When exposed to light, the muconate anions polymerize to generate molecule-thick sheets of polymers (green) between the alkylammonium ions. **c**, These sheets can be liberated with acid (purple), but can then be restacked with a molecular adhesive, consisting of a different alkylammonium ion (red chains) to generate a new material with identical sheets (green) but with different layer-to-layer spacing (d_2).

of hydrogen bonding in sheet formation from what are essentially one-dimensional polymer chains. Do the sheets actually remain intact following exposure to acid, perhaps because of hydrogen bonding between the polymer chains, or are the chains dispersed by the acid and then reassembled into sheets when the molecular glue is added? The precise arrangement and structure of the alkylammonium ions on the surface of the polymer sheets is an open question, as it is with most other intercalates in layered systems¹⁵.

The generality of this new approach is yet to be demonstrated, but it might be possible to use derivatives of muconic acid to build similar crystalline sandwiches. Using the power of organic synthesis, it should be feasible to create new topochemically reactive materials based on such derivatives, perhaps with functional groups that introduce useful electronic or optical properties¹⁶. The work by Matsumoto *et al.* is an important first step in that direction.

Michael D. Ward is in the Department of Chemical Engineering, University of Minnesota, 151

Amundson Hall, 421 Washington Avenue SE, Minneapolis 55455-0132, USA.

e-mail: wardx004@maroon.tc.umn.edu

1. Matsumoto, A., Odani, T., Sada, K., Miyata, M. & Tashiro, K. *Nature* **405**, 328–330 (2000).
2. Cohen, M. D. & Schmidt, G. M. J. *J. Chem. Soc.* 1996–2000 (1964).
3. Schmidt, G. M. J. *Pure Appl. Chem.* **27**, 687–688 (1971).
4. Desiraju, G. R. *Crystal Engineering: the Design of Organic Solids* (Elsevier, Amsterdam, 1989).
5. Cheng, K. & Foxman, B. M. *J. Am. Chem. Soc.* **99**, 8102–8103 (1977).
6. Cao, G. & Mallouk, T. E. *J. Solid State Chem.* **94**, 59–71 (1991).
7. Tieke, B. & Wegner, G. *Angew. Chem. Int. Edn Engl.* **20**, 687 (1981).
8. Day, P. & Ledsham, R. D. *Mol. Cryst. Liq. Cryst.* **86**, 1903–1914 (1982).
9. Enkelmann, V. *Adv. Polym. Sci.* **63**, 91–136 (1984).
10. Kane, J. J., Liao, R.-F., Lauher, J. W. & Fowler, F. W. *J. Am. Chem. Soc.* **117**, 12003–12004 (1995).
11. Fogg, A. M., Green, V. M., Harvey, H. G. & O'Hare, D. *Adv. Mater.* **11**, 1466–1469 (1999).
12. Kaschak, D. M. *et al. J. Am. Chem. Soc.* **120**, 10887–10894 (1998).
13. Kovtyukhova, N. I. *et al. Chem. Mater.* **11**, 771–778 (1999).
14. Kagan, C. R., Mitzi, D. B. & Dimitrakopoulos, C. D. *Science* **286**, 945–947 (1999).
15. Cai, H., Hillier, A. C., Franklin, K., Nunn, C. C. & Ward, M. D. *Science* **266**, 1551–1555 (1994).
16. Mitzi, D. B., Chondroudis, K. & Kagan, C. R. *Inorg. Chem.* **38**, 6246–6256 (1999).

Daedalus

Drink without drinking

The liver is a vigilant guardian of our health. All the blood from the stomach and digestive tract passes through it before entering general circulation. Any poisons or dangerous substances we have swallowed are likely to be detoxified by the liver before they can do any damage. This prudent arrangement annoys pharmacologists, many of whose drugs are badly degraded by the liver before they can reach the organ in need of them. It also sabotages drinkers, for the liver rapidly oxidizes alcohol. So Daedalus is looking for some way of by-passing the liver. Dedicated alcoholics could then get drunk on far less alcohol.

Direct injection seems an unattractive option. Skin-absorption seems better, yet few drinkers would want to sit in a bath of beer. But Daedalus recalls that the skin is very permeable to fatty substances. Suppose the alcohol was converted into a fatty ester such as ethyl oleate, and smeared on the skin. It would be absorbed rapidly. Once inside, the body's esterases would hydrolyse it to alcohol. The challenge is to smear the ester on a site whose subsurface blood flows, not towards the liver, but to the brain. Somewhere on the neck, over the carotid artery, might perhaps work. The alcohol would then reach the target organ directly, without wastefully saturating all the rest of the tissues. A very small quantity, easily within smearing volume, would suffice; and the liver would be helpless to intervene.

So would the taxman. Ethyl oleate, not being alcohol, would be free of excise duty. Even if the law caught up with the trick and deemed ethyl esters to be potable alcohol, the duty on such a small amount would be trivial.

Even the excuse for excise duty, the social damage caused by alcohol, would largely vanish. Daedalus' 'Rubbing Alcohol' will give the user total control of his habit. With such small quantities in play, he will reach and maintain his desired state of intoxication very rapidly. And when the ester has been absorbed and metabolized, he will recover sobriety equally fast. After a cheerful evening's ester carouse, he could sober up in minutes and drive home entirely safely. He might, of course, wish to maintain tradition by drinking large amounts of alcohol-free wine or beer at the same time. His liver would be relieved of its burden; his bladder would not.

David Jones

The Further Inventions of Daedalus is published by Oxford University Press.

Guiding the swing in golf putting

Golfers control the pace of a putt by comparing sensory data with an internal guide.

Actions that involve making contact with surfaces often demand perceptual regulation of the impact — for example, of feet with ground when walking or of bat with ball when hitting. Here we investigate how this control of impact is achieved in golf putting, where control of the club-head motion at ball impact is paramount in ensuring that the ball will travel the required distance. Our results from ten professional golfers indicate that the clubhead motion is spatially scaled and perceptually regulated by coupling it onto an intrinsic guide generated in the nervous system.

Our model of motor control in putting considers two fundamental but largely ignored issues^{1,2}: the role that information gathered through the senses plays in guiding actions, and the control processes used by the nervous system to solve the guidance problem. This model (Box 1, overleaf) is based on a general theory^{3–5} that any guiding movement must involve coordinating and regulating the rate of closure of motion gaps, such as the gap between a club and ball.

The theory argues that gap closure is controlled by using perceptual information about a particular measure of a gap, denoted as τ , which is equal to the time to gap closure at the existing closure rate. The rate of gap closure can be regulated by keeping the τ value of the motion gap coupled onto (in constant ratio with) the τ value of a 'guiding gap'. In self-paced actions like putting, this τ guide is assumed to be generated in the nervous system (for example, by modulating energy levels). Neural mechanisms are implicated in the intrinsic timing of movements^{6–8} and in their spatial specification⁹, and there are similarities in brain activity when the performance of an action is being imagined and when it is subsequently executed^{10–12}.

The golfer controls the motion pattern of the forward swing of the putting action by constantly sensing the τ of the gap between the club and the end of the follow-through, and keeping this τ in constant ratio with an intrinsic τ guide (Box 1, Fig. 1a)^{4,5}. Spatially scaling the forward-swing pattern generated by the intrinsic τ guide could be achieved by adjusting one or more of four possible parameters (Box 1): the amplitude, D , of the forward swing (more specifically, D^2); the duration, T , of the forward swing (more specifically, $1/T^2$); the relative time, P_T , at which the ball is hit; and the τ -coupling constant, k . The first two parameters are the most plausible as they alone are linearly related to the distance the ball will travel (equation (2) in Box 1).

To test this model, we analysed the putting actions of ten low-handicap (under 5) golfers across different distances. The motion pattern was τ -guided, as evidenced by the strong linear relation between the temporal evolution of the τ of the forward swing and the intrinsic τ guide (Fig. 1b). The proportion of energy loss at club-ball impact and the retarding force on the ball were both relatively constant, as assumed by the model, with r^2 linear-regression values of the club's impact velocity (squared) onto

putting distance being in the range 0.985 to 0.999. Spatial scaling was brought about by adjusting D^2 , whereas the other three parameters showed less systematic change with putting distance (Fig. 1c–f).

Our model predicts how golfers regulate the spatial and temporal components of the forward swing in order to transmit the appropriate amount of kinetic energy at ball impact. The same principles — of guiding movements by coupling them onto intrinsic τ guides and spatially scaling the

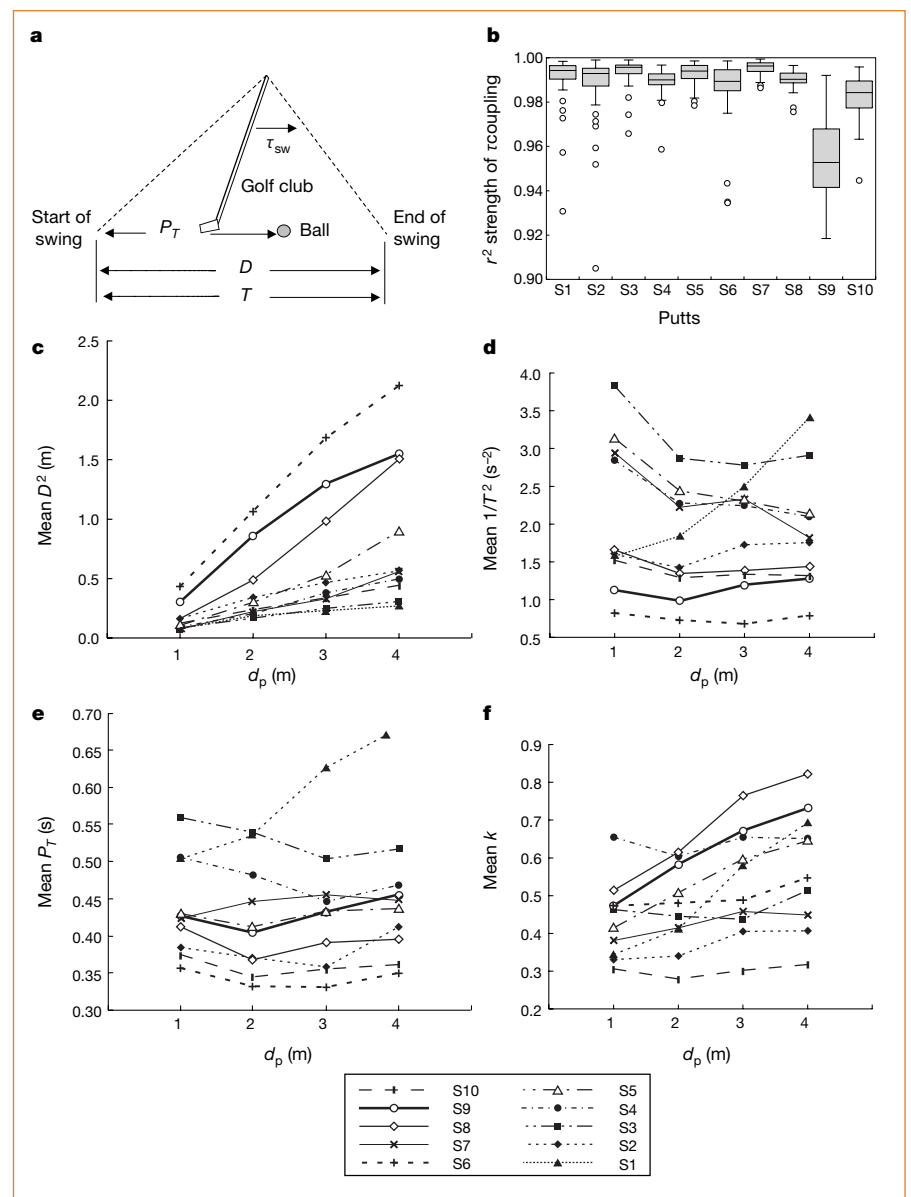


Figure 1 τ -guiding the forward swing. Ten golfers' club movements were recorded at 200 Hz when putting to four distances, ten times, on an artificial green. **a**, Motion parameters defined in Box 1. Horizontal distance/velocity ratio for the clubhead during the forward swing was a measure of the τ_{sw} . For each putt, τ_{sw} was linearly regressed onto the intrinsic τ guide (τ_g) to determine the degree of linearity (r^2). **b**, Range of the proportions of variance accounted for by the τ -coupling model (as measured by r^2 values) for golfers S1–S10. **c–f**, Relation between the parameters (D^2 , $1/T^2$, P_T and k) and putt distance, d_p , for each of the ten golfers.

Box 1 The golfer's τ guide to putting

The golfer guides the forward swing by coupling τ_{sw} (time to gap closure, at the prevailing closure rate of the club-head with the end of the follow-through) onto an intrinsic τ -guide, $\tau_g = 0.5(t - T^2/t)$, where t is the time from the start of the swing and T is its duration. Thus, the relation $\tau_{sw} = k\tau_g = 0.5k(t - T^2/t)$ is maintained during the forward swing (k is a coupling con-

stant). Integration leads to $V_c = 2D(1/T)(P_r/K)(1 - P_r^2)^{1/8-1}$ (1) where V_c is the clubhead velocity just before impact, D is the amplitude of the forward swing and P_r is the proportion of the swing duration before the ball is hit. The kinetic energy of the club-body system (effective mass, M_c) just before impact is equal to $M_c V_c^2/2$. Assuming a constant proportion of energy is lost

during impact and the ball's motion is opposed by a constant force F , equation (2) gives the distance putted (d_p) as $d_p = \lambda M_c V_c^2 / 2F = (2\lambda M_c / F) D^2 (1/T^2) (P_r/K)^2 (1 - P_r^2)^{2/8-2}$, where λ is a constant. Thus, the golfer could scale the distance putted by adjusting D^2 , $1/T^2$, P_r or k . As D^2 and $1/T^2$ are each linearly related to distance, these are the more likely candidates.

movements — might apply generally when regulating impact force with objects and surfaces. The parameter used in scaling, however, might vary with the task.

C. M. Craig*, D. Delay†, M. A. Grealy‡, D. N. Lees§

*UMR Mouvement et Perception, 163 Avenue de Luminy, Université de la Méditerranée, 13288 Marseilles, France

e-mail: craig@laps.univ-mrs.fr

†UFRAPS, Université de Joseph Fourier, 38041 Grenoble, France

‡Department of Psychology, Strathclyde University, Graham Hills Building, Glasgow G1 1XQ, UK

§Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK

1. Kelso, J. A. S. *Hum. Mov. Sci.* **10**, 93–111 (1991).
2. Hudson, J. L. *Med. Sci. Sport Ex.* **18**, 242–251 (1986).
3. Lee, D. N. *Ecol. Psychol.* **10**, 221–250 (1998).
4. Craig, C. M. & Lee, D. N. *Exp. Brain Res.* **124**, 371–382 (1999).
5. Lee, D. N., Craig, C. M. & Grealy, M. A. *Proc. R. Soc. Lond. B* **266**, 2029–2035 (1999).
6. Ivry, R. & Keele, S. J. *Cogn. Neurosci.* **1**, 136–152 (1999).
7. Nichelli, P., Alway, D. & Grafman, J. *Neuropsychology* **34**, 863–871 (1996).
8. Meck, W. H. *Cogn. Brain Res.* **3**, 227–242 (1996).
9. Georgopoulos, A. P. & Massey, J. T. *Exp. Brain Res.* **65**, 361–370 (1987).
10. Jeannerod, M. *Neuropsychology* **33**, 1419–1432 (1995).
11. Jeannerod, M. *Behav. Brain Sci.* **17**, 187–245 (1999).
12. Georgopoulos, A. P. et al. *Science* **243**, 234–236 (1989).

Natural selection

Evolution of lifespan in *C. elegans*

It was proposed almost 50 years ago that ageing is non-adaptive and is the consequence of a decline in the force of natural selection with age¹. This led to the theory that ageing results from detrimental effects late in life of genes that act beneficially in early life^{1,2}, so any genetic alteration that increases lifespan might be expected to reduce fitness, for example. We show here that a mutation that greatly increases the lifespan of the nematode *Caenorhabditis elegans* does indeed exhibit a fitness cost, as demonstrated during starvation cycles that may mimic field conditions, thereby validating the pleiotropy theory of ageing².

C. elegans is a soil-dwelling nematode with a facultative, self-fertilizing mode of reproduction. Mutation of the *age-1* gene, which encodes a phosphatidylinositol 3-OH kinase catalytic subunit component of the insulin-like signalling pathway³, can extend adult lifespan by up to 80% (ref. 4). The *age-1* gene, and other genes encoding components of the insulin-like signalling pathway, not only influence ageing, but also control progress of normal development^{5–7} and determine adult stress resistance⁸.

If worms carrying the weak mutant allele *age-1(hx546)* are grown at 27 °C, they develop into dauer larvae (a diapause stage) that does not feed or reproduce³, but if they are grown at 20 °C, they develop normally

into adults. At this growth temperature, mutant and wild-type worms are essentially identical in appearance, development rates, activity levels and total fertility^{4,9,10}.

Alleles that confer an extended lifespan but appear otherwise normal are at odds with the pleiotropy theory^{1,2}. We have tested this theory by measuring the relative fitness of the *hx546* and wild-type alleles of *age-1* in strains that are otherwise isogenic.

We established synchronously ageing populations of hermaphrodite worms, each containing wild-type and *age-1(hx546)* worms on the same agar plates at 20 °C. As there were no males in these populations, all

progeny are the result of self-fertilization. To determine the allele frequency at the *age-1* locus in each generation, 100 eggs were removed and shifted to 27 °C, where after 3 days wild-type worms developed into adults and *age-1(hx546)* worms developed into dauer larvae. The ratio of adults to dauers thus indicated the *age-1* allele frequency in the populations. Meanwhile, the populations were maintained by transferring eggs to new plates at 20 °C. When the populations were maintained over 10 generations on agar plates spotted with *Escherichia coli* as a constant food source, there was no consistent change in allele frequency (Fig. 1a), even when the *hx546* allele frequency in the founding populations was varied from 0.1 to 0.9. Thus, there is no evidence of a trade-off between longevity and other life-history traits under these conditions.

This was not the case when mixed populations were maintained under conditions of cyclical starvation. We allowed the worms to exhaust the bacterial food and starved them for four days. Only eggs laid within 24 hours were used to initiate the next cycle. After six starvation cycles, a homogeneous response ($G = 7.20$, d.f. = 4, $P = 0.13$) was observed in all populations, with *hx546* changing from an initial frequency of 0.50 to a mean of 0.06 (Fig. 1b). Such a large reduction in allele frequency suggests a substantial difference in relative fitness under starvation conditions.

To test which life-cycle stage contributed to the change in allele frequency, we picked starved worms from population plates, allowed them to feed and examined them 12 and 24 h later. Only young adults laid eggs over this period, indicating that the trade-off occurs during this life-cycle stage.

We conclude that the extension of lifespan, by mutation of a single gene, is associated with reduced fitness. This fitness cost is only apparent in an environment thought

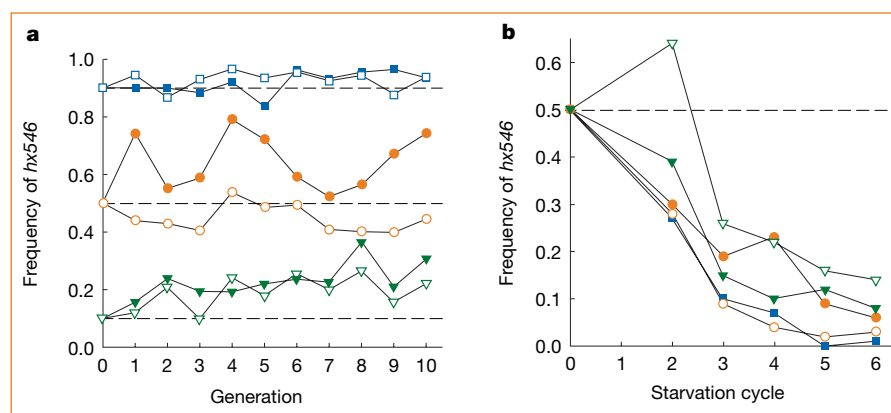


Figure 1 Direct competition between long-lived and wild-type worms in laboratory natural-selection experiments. **a**, Frequencies of *age-1(hx546)* under non-starvation conditions at 20 °C in mixed populations of N2 (wild type) and TJ1052 (*age-1(hx546)*) worms. **b**, Frequencies of *age-1(hx546)* in five replicate populations cultured with cyclical starvation, where the worms were depleted of all food by five days and remained starved for a further four days. Each starvation cycle approximated two generations. For a reduction in allele frequency of 0.44 in 12 generations, the relative fitness of *age-1(hx546)* versus wild type is estimated to be 0.77, indicating a major cost to *age-1(hx546)* under these conditions.

to mimic natural conditions. However, we have also demonstrated that, in the presence of food, genetic lifespan extension can be uncoupled from other life-history traits to the extent that fitness is not affected.

The observed fitness cost associated with field-like conditions provides an example of a single gene that acts in early life, ageing and Darwinian fitness, as predicted by the pleiotropy theory of ageing².

David W. Walker, Gawain McColl,

Nicole L. Jenkins, Jennifer Harris,

Gordon J. Lithgow

School of Biological Sciences, University of Manchester, Manchester M13 9PT, UK

e-mail: Gordon.Lithgow@man.ac.uk

1. Medawar, P. B. *An Unsolved Problem of Biology* (Lewis, London, 1952).
2. Williams, G. C. *Evolution* **11**, 398–411 (1957).
3. Morris, J. Z., Tissenbaum, H. A. & Ruvkun, G. *Nature* **382**, 536–539 (1996).
4. Friedman, D. B. & Johnson, T. E. *J. Gerontol.* **43**, B102–B109 (1988).
5. Malone, E. A., Inoue, T. & Thomas, J. H. *Genetics* **143**, 1193–1205 (1996).
6. Tissenbaum, H. A. & Ruvkun, G. *Genetics* **148**, 703–717 (1998).
7. Kenyon, C. et al. *Nature* **366**, 461–464 (1993).
8. Lithgow, G. J., White, T. M., Melov, S. & Johnson, T. E. *Proc. Natl Acad. Sci. USA* **92**, 7540–7544 (1995).
9. Duhon, S. A. & Johnson, T. E. *J. Gerontol. Biol. Sci. Med. Sci.* **50**, B254–B261 (1995).
10. Lithgow, G. J., White, T. M., Hinerfeld, D. A. & Johnson, T. E. *J. Gerontol.* **49**, B270–B276 (1994).

Auditory perception

How cicadas interpret acoustic signals

The vertebrate ear can analyse the frequency components of sound with high resolution, recognizing complex acoustic signals even against a noisy background¹. By contrast, insect ears can separate only broad-frequency bands, resulting in a categorical perception of sound². We have discovered an insect, the cicada *Tettigetta josei*, that has a capacity for fine-frequency resolution, which could explain the evolution of frequency-modulated communication signals in cicadas.

Although known insects can discriminate only broad-frequency bands, their individual receptors can often respond selectively to frequencies within such a band and so might allow the central nervous system to perform fine-scale frequency analysis^{3,4}. But this is not observed: receptor information is pooled either for distance estimation of an acoustic signal (exploiting the frequency-dependent attenuation of sound travelling in air⁵) or for sharpening the frequency band employed in acoustic communication by lateral inhibition⁶.

Most hearing-insect groups have only a few auditory receptors in their tympanal organs (2–200; Table 1), which may constrain sound analysis in the frequency and

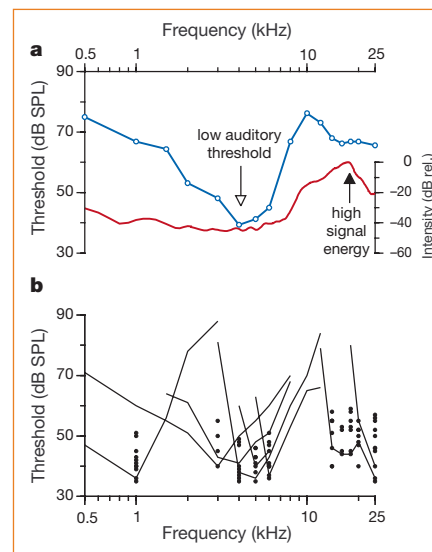


Figure 1 Song spectrum of the cicada *Tettigetta josei*, sensitivity of its auditory system, and sharp tuning of its interneurons to different frequencies. **a**, Threshold curve from extracellular auditory nerve recordings (circles; mean from $n=9$) and spectrum of the song signal (average recorded from 8 males). The lowest threshold of this cicada's ear lies at 3–6 kHz (open arrow) and is not matched by the high-energy peak of the male signal at higher frequencies (filled arrow). **b**, Tuning curves from intracellular recordings made from 8 interneurons with different characteristic frequencies. Each point below 10 kHz represents the absolute threshold of an interneuron at its characteristic frequency. At frequencies above 10 kHz, some points stem from the same cell. Whereas recordings from the whole tympanal nerve reflect the threshold of the majority of auditory receptors (**a**), the differently tuned receptors remain undetected; however, these are revealed by sensitive interneurons at 1–1.5 kHz and 15–25 kHz (**b**). Full details of the methods are available as Supplementary Information.

time domains¹. By contrast, cicadas have up to 2,100 auditory receptors in each ear, rivaling or even exceeding the number in most lower vertebrates (Table 1).

Extracellular recordings of summed potentials of auditory receptors suggest that cicadas have a uniform tuning to a low-frequency band (3–6 kHz^{7,8}), implying that for many of these insects the spectral energy of the relevant communication signals does not match the frequency range of best hearing (in 17 out of 22 species^{7,8}). But differential tuning of these receptors in a cicada's ear might not only explain the apparent and — if true — maladaptive mismatch of the communication system, but also identify a function for the many receptors involved in frequency discrimination.

To test this idea of differential receptor tuning, we recorded intracellularly from auditory interneurons in the Portuguese cicada *T. josei*, which has a large number of receptors as well as a mismatch of frequency bands between the spectrum of the emitted signal and the apparent tuning of the audi-

tory system (Fig. 1a). A set of interneurons tuned sharply to different frequencies shows that the thoracic nervous system is capable of frequency discrimination (Fig. 1b). At these frequencies, interneurons are extremely sensitive (thresholds of 30–50 dB sound pressure level) and many cells show 'roll-offs' of 30–50 dB per octave to one or both sides of their characteristic frequencies (a roll-off is a decrease in the sensitivity of auditory interneurons to adjacent frequencies, in octaves, to their characteristic frequency). These roll-offs convert to $Q_{10\text{ dB}}$ values from 2.8 to 7.4 (Table 1, legend) which — for insects — are remarkably high and are similar to $Q_{10\text{ dB}}$ values found for most lower vertebrates.

We conclude that these interneurons provide this cicada with a much lower threshold for intraspecific communication signals than would be predicted from recordings of the summed activity of auditory receptors, thereby solving the enigmatic frequency mismatch in the communication system of these insects (Fig. 1a). At

Table 1 Comparison of the tympanal ears of vertebrates and insects

	Number of auditory receptors	$Q_{10\text{ dB}}$	Frequency range (kHz)
Vertebrates			
Fish	> 5,000	0.35–1 (2)	0.05–1
Amphibians	14–1,500	0.3–3	0.1–4
Reptiles	50–2,000 (12,000)	1–7	0.01–5
Birds	5,800–9,600	2–12	1–10
Mammals	2,500–14,000	2–200	0.1–200
Insects			
Moth	2–4	<1	10–120
Mantid	20–35	<1	20–100
Bushcricket	20–60	1–2	0.1–80
Cricket	50–70	1–2	0.1–60
Grasshopper	30–80 (2,000)	~1	0.1–50
Fly	30–200	<1	1–40
Cicada	600–2,100	1–7	0.1–25

The number of auditory receptors in the ears of vertebrates (hair cells) and insects (scolopodial cells) is shown, together with the $Q_{10\text{ dB}}$ values and frequency range within their auditory pathways. $Q_{10\text{ dB}}$ values provide a measure of frequency tuning and are derived from a neuron's tuning curve at 10 dB above the threshold of the characteristic frequency (CF, the frequency at which a cell shows the lowest threshold); $Q_{10\text{ dB}}$ is defined by CF divided by the bandwidth at 10 dB above the threshold of CF. The ranges given correspond to those commonly observed between species in a given group; numbers in parentheses are exceptions. References for the data shown are available as Supplementary Information.

frequencies of 1–6 kHz and 12–25 kHz, thresholds between cells with distinct characteristic frequencies differ by 20–30 dB at a given frequency (Fig. 1b), thus providing a unique facility (for insects) for frequency discrimination in the auditory pathway.

Given the uniform design of the auditory system in cicadas⁷, it is possible that the frequency-modulated songs of many cicada species, particularly tropical ones⁹, may result from sensory drive¹⁰, because females are able to use frequency components of songs as criteria for species recognition and mate choice.

P. J. Fonseca*, D. Münch†, R. M. Hennig†

*Departamento Zoologia and Centro de Biologia Ambiental, Faculdade Ciências Lisboa, Campo Grande, 1700 Lisboa, Portugal

†Institut für Biologie, Humboldt Universität zu Berlin, 10 115 Berlin, Germany

e-mail: matthias.hennig@rz.hu-berlin.de

1. Bradbury, J. W. & Vehrencamp, S. L. *Principles of Animal Communication* (Sinauer, Sunderland, MA, 1998).
2. Hoy, R. R., Popper, A. N. & Fay, R. R. (eds) *Comparative Hearing: Insects* (Springer, Berlin, 1998).
3. Römer, H. *Nature* **306**, 60–62 (1983).
4. Oldfield, B. P. *Trends Neurosci.* **11**, 267–270 (1988).
5. Römer, H. *Phil. Trans. R. Soc. Lond. B* **226**, 179–185 (1993).
6. Stumpner, A. J. *Neurophysiol.* **79**, 2408–2415 (1998).
7. Fonseca, P. J. *Acoustic Communication in Cicadas (Homoptera, Cicadoidea): Sound Production and Sound Reception*. Thesis, Univ. Lisbon (1994).
8. Popov, A. V. in *Sensory Systems and Communication in Arthropods* (eds Gribakin, F. G., Wiese, K. & Popov, A. V.) 301–304 (Birkhäuser, Basel, 1990).
9. Gogala, M. *Bioacoustics* **6**, 101–116 (1995).
10. Endler, J. A. *Am. Nat.* **139**, 125–153 (1992).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Biotechniques

Transfection of cells by immunoporation

Cell transfection is now a central technique in molecular biology and an essential prerequisite for gene therapy. Here we describe how beads coated with antibodies and bound to specific cell-surface transmembrane proteins can create holes in cells when the beads are removed, allowing transfection of the cells with DNA or other macromolecules. This unique targeted transfection of cells by immunoporation is very efficient and results in minimal cell death.

A variety of methods have been developed for the transfection of cells, including electroporation^{1,2}, lipofection^{3,4}, calcium phosphate coprecipitation^{5,6} and DEAE dextran^{7,8}. Of these methods, only electroporation offers the possibility of introducing DNA and other molecules such as proteins into viable cells. None of the current methods is able to target specific types of cells for transfection. In this new method

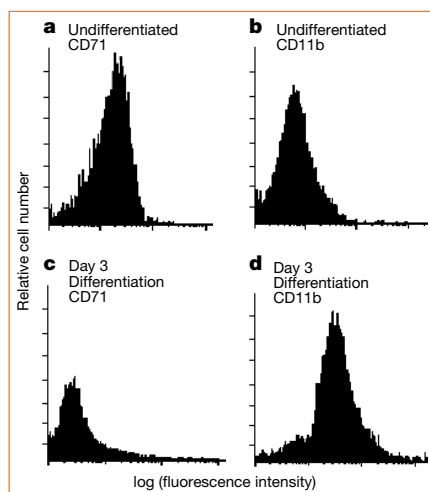


Figure 1 Analysis of the transfection of HL-60 cells with pEGFP-C1 by measurement of the expression of green fluorescent protein using flow cytometry. **a**, HL-60 cells transfected using DYNAFECT-CD71 beads; **b**, HL-60 cells transfected using DYNAFECT-CD11b beads; **c**, DMSO-induced HL-60 cells transfected using DYNAFECT-CD71 beads; **d**, DMSO-induced HL-60 cells transfected using DYNAFECT-CD11b beads.

of cell transfection, antibody-coated beads are bound to specific surface antigens and then the beads are sheared off from the cell by mixing: this causes the formation of transient holes in the cell membrane through which macromolecules can enter.

Granulocytic, differentiating human lymphoblastic HL-60 cells normally express CD71 on their surfaces. When induced to differentiate in the presence of dimethyl sulphoxide (DMSO), the cells cease to express CD71 and instead express CD11b. We have used this cell line as a model system to investigate the process of cell transfection mediated by immunoporation.

DYNAFECT beads coated with either anti-CD11b antibody (DYNAFECT-CD11b) or anti-CD71 antibody (DYNAFECT-CD71) were mixed on a rotating end-over-end mixer at 33 r.p.m. for 6 h at 22 °C with either uninduced HL-60 cells or cells that had been induced with DMSO for 3 days. For mixing in a 2-ml microfuge tube, 10⁷ beads and 5 × 10⁵ cells were suspended in 0.5 ml transfection medium (Dyna AS) containing 0.2 µg plasmid DNA vector pEGFP-C1 (4.7 kilobases) which codes for green fluorescent protein. After transfection, the beads were removed using a magnetic separator and the cells were transferred back into tissue-culture medium and cultured for a further 48 hours before analysis.

The extent of cell transfection was determined by flow cytometry. Figure 1 shows that the DYNAFECT-CD71 beads facilitate the transfection of DNA into normal HL-60 cells, but when the cells become differentiated and no longer express CD71, transfection no longer occurs with these beads. In contrast, mixing undifferentiated HL-60

cells with DYNAFECT-CD11b beads does not result in the transfection of the cells with DNA, but when the cells are differentiated and begin to express CD11b, those beads do bring about transfection.

Hence, immunoporation has the potential to target specific types of cell in a mixed population for transfection, depending on their immunological identity, and allow the targeted cells to take up a variety of different molecules. This also occurs in several mammalian cell lines with a range of different antibodies that target selected cell-surface antigens. In all cases, the level of transfection was 40–80%, depending on mixing conditions, and non-viable cells usually numbered less than 20%.

The high levels of selectivity and transfection, together with minimal cell death, that are achievable with immunoporation illustrate the enormous potential of this technique for use in a wide range of transfection studies. In particular, the ability to target specific subpopulations of cells will be extremely useful for many gene therapy applications.

Lale Bildirici, Patricia Smith, Christos Tzavelas, Elina Horefti, David Rickwood

Department of Biology, University of Essex, Colchester CO4 3SQ, UK

e-mail: rickd@essex.ac.uk

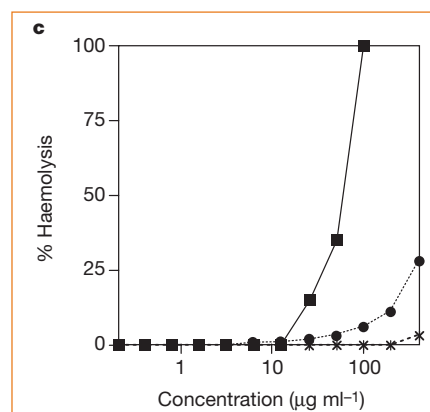
1. Fromm, M., Callis, J., Taylor, L. P. & Walbot, V. *Methods Enzymol.* **153**, 351 (1987).
2. Pfau, J. & Youderian, P. *Nucleic Acids Res.* **18**, 6165 (1990).
3. Watanabe, Y. et al. *J. Biochem.* **116**, 1220–1226 (1994).
4. Hargest, R., Eldin, A. & Williamson, R. *Adv. Exp. Med. Biol.* **451**, 385–391 (1998).
5. Ambrosini, E. et al. *J. Neurosci. Res.* **55**, 569–577 (1999).
6. Nielson, T. et al. *Mol. Gen. Genet.* **242**, 280–288 (1994).
7. Muck, K. D., Wei, R. & Elbagarri, A. *J. Immunol. Methods* **211**, 79–86 (1998).
8. Yang, Y. W. & Yang, J. C. *Biotechnol. Appl. Biochem.* **25**, 47–51 (1997).

Erratum

Non-haemolytic β-amino-acid oligomers

E. A. Porter, X. Wang, H.-S. Lee, B. Weisblum, S. H. Gellman *Nature* **404**, 565 (2000)

Some symbols representing haemolytic activity were absent or incorrect in Fig. 1c. The correct figure is shown here. Crosses, β-17; circles, magainin derivative; squares, melittin.



Lateral gene transfer and the nature of bacterial innovation

Howard Ochman*, Jeffrey G. Lawrence† & Eduardo A. Groisman‡

* Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721-0088, USA

† Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA

‡ Howard Hughes Medical Institute, Washington University School of Medicine, Department of Molecular Microbiology, St Louis, Missouri 63110, USA

Unlike eukaryotes, which evolve principally through the modification of existing genetic information, bacteria have obtained a significant proportion of their genetic diversity through the acquisition of sequences from distantly related organisms. Horizontal gene transfer produces extremely dynamic genomes in which substantial amounts of DNA are introduced into and deleted from the chromosome. These lateral transfers have effectively changed the ecological and pathogenic character of bacterial species.

Given that they are single-celled organisms and that their genome sizes vary by little more than an order of magnitude in length, bacteria display extraordinary variation in their metabolic properties, cellular structures and lifestyles. Even within relatively narrow taxonomic groups, such as the enteric bacteria, the phenotypic diversity among species is remarkable. Although the enteric bacteria as a whole share a myriad of traits denoting their common ancestry, each species also possesses a unique set of physiological characteristics that define its particular ecological niche (Fig. 1).

Several mechanisms could be responsible for the differences evident among bacterial species. Point mutations leading to the modification, inactivation or differential regulation of existing genes have certainly contributed to the diversification of microorganisms on an evolutionary timescale; however, it is difficult to account for the ability of bacteria to exploit new environments by the accumulation of point mutations alone. In fact, none of the phenotypic traits that are typically used to distinguish the enteric bacteria *Escherichia coli* from its pathogenic sister species *Salmonella enterica* can be attributed to the point mutational evolution of genes common to both¹. Instead, there is growing evidence that lateral gene transfer has played an integral role in the evolution of bacterial genomes, and in the diversification and speciation of the enterics and other bacteria.

The significance of lateral gene transfer for bacterial evolution was not recognized until the 1950s, when multidrug resistance patterns emerged on a worldwide scale². The facility with which certain bacteria developed resistance to the same spectrum of antibiotics indicated that these traits were being transferred among taxa, rather than being generated *de novo* by each lineage. Although the widespread impact of lateral gene transfer on bacterial evolution was not appreciated until much later, these early studies of rapid evolution by gene acquisition encompass four issues relevant to all current studies. First, how is it possible to detect and to identify cases of lateral gene transfer? Second, where do these genes come from, and by what mechanisms are they transferred? Third, what types of, and how many, traits have been introduced through lateral gene transfer? And fourth, at what relative rates are different classes of genes mobilized among genomes?

Detecting lateral gene transfer

How is it possible to establish whether a new trait, or a specific genetic region, is the result of horizontal processes? Naturally, it would be most satisfying to actually observe the conversion of a deficient strain in the presence of an appropriate donor—and all the more convincing to establish that genetic material had indeed been transferred and the manner in which it was acquired. But outside experimental settings—that is, for most cases of lateral gene transfer

in the recent or evolutionary history of a bacterial species—actual transfer events are only rarely observed, and unambiguous evidence of their occurrence must be derived from other sources.

Lateral gene transfer creates an unusually high degree of similarity between the donor and the recipient strains for the character in question. Furthermore, because each transfer event involves the introduction of DNA into a single lineage, the acquired trait will be limited to the descendants of the recipient strain and absent from closely related taxa, thereby producing a scattered phylogenetic distribution. However, lateral gene transfer need not be invoked to explain the sporadic occurrence of certain phenotypic traits, such as the ability to withstand particular antibiotics, because these properties can originate through point mutations in existing genes and therefore may evolve independently in divergent lineages. Thus, additional information is needed to discriminate between convergent evolution and lateral gene transfer. Clearly, the strongest evidence for (or against) lateral gene transfer derives from a molecular genetic analysis of the DNA sequences themselves.

DNA sequence information has been used in diverse ways to identify cases of lateral gene transfer, but the underlying basis of most applications is to discover features indicating that the evolutionary history of genes within a particular region differs from that of ancestral (vertically transmitted) genes. Similar to distinctive phenotypic properties, DNA segments gained through lateral gene transfer often display a restricted phylogenetic distribution among

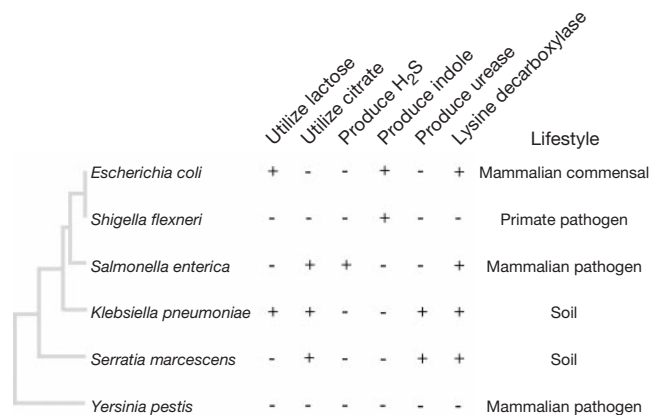


Figure 1 Evolutionary relationships and phenotypic profiles of representative enteric bacteria. Plus denotes the presence and minus the absence of a trait in more than 85% of strains. Evolutionary relationships among species are based on nucleotide sequence information. In many cases, genes acquired by horizontal transfer confer the species-specific traits.

related strains or species. In addition, these species-specific regions may show unduly high levels of DNA or protein sequence similarity to genes from taxa inferred to be very divergent by other criteria³. The significance of aberrant phylogenies can be evaluated by phylogenetic congruency tests or other means⁴.

Although gene comparisons and their phylogenetic distributions are useful for detecting lateral transfer, the DNA sequences of genes themselves provide the best clues to their origin and ancestry within a genome. Bacterial species display a wide degree of variation in their overall G+C content, but the genes in a particular species' genome are fairly similar with respect to their base compositions, patterns of codon usage and frequencies of di- and trinucleotides⁵⁻⁷. Consequently, sequences that are new to a bacterial genome, in other words, those introduced through horizontal transfer, retain the sequence characteristics of the donor genome and thus can be distinguished from ancestral DNA¹.

It is not surprising that genomic regions often manifest several attributes that denote their acquisition through lateral gene transfer. For example, a large number of *S. enterica* genes that are not present in *E. coli* (or any other enteric species) have base compositions that differ significantly from the overall 52% G+C content of the entire chromosome^{8,9}. Within *S. enterica*, certain serovars (that is, lineages that exhibit a distinct composition of flagellar and/or lipopolysaccharide surface antigens) may contain more than a megabase of DNA not present in other serovars, as assessed by a genomic subtraction procedure. The base compositions of these anonymous serovar-specific sequences suggest that at least half were gained through horizontal transfer¹⁰. In addition to information obtained from the sequences of the genes themselves, the regions adjacent to genes identified as being horizontally transferred often contain vestiges of the sequences affecting their integration, such as remnants of translocatable elements, transfer origins of plasmids or known attachment sites of phage integrases, which further attest to their foreign origin in the genome.

Genetic exchange within and between bacterial species also acts upon homologous sequences, and numerous techniques have been developed to detect such events from sequence data¹¹. However, the action of mismatch correction systems greatly reduces the efficiency of homologous recombination when donor and recipient sequences contain nonidentical bases^{12,13}. As a result, homologous recombination is most successful in integrating DNA into the chromosome when the donor and recipient are relatively closely related. Because this type of genetic exchange principally affects the variation in existing genes, rather than introducing unique traits to the genome, its role in the ecological and physiological diversification of bacteria is apt to be negligible.

The scope of lateral gene transfer

The analysis of individual genes has uncovered numerous cases of lateral gene transfer; however, the ability to recognize horizontally acquired regions on the basis of their sequence characteristics makes it possible to assess the total proportion of foreign genes within a genome without resorting to gene phylogenies, sequence alignments or homology searches. Early attempts to establish the extent of laterally transferred sequences in a genome were limited to the very few microorganisms for which there was sufficient sequence information to get an unbiased sample of genes. Using this approach, it was originally estimated that between 10% and 16% of the *E. coli* chromosome arose through lateral gene transfer^{9,14,15}—a range similar to the amount of unique DNA in the *E. coli* chromosome inferred from early alignments of the *E. coli* and *S. enterica* genetic maps¹⁶.

The availability of complete genomic sequences provides an opportunity to measure and compare the cumulative amount of laterally transferred sequences in diverse bacterial genomes (Fig. 2). Potentially foreign genes are identified by their atypical nucleotide compositions, or patterns of codon usage bias^{1,9}; after correcting for genes whose atypical features are due to amino-acid composition, the remaining genes are likely to have been introduced relatively recently by lateral gene transfer. Wide variation has been observed in the size and organization of 19 genomes analysed, and the amount of horizontally acquired DNA—represented as those open reading frames (ORFs) whose sequence characteristics depart from the prevalent features of their resident genome—ranges from virtually none in some organisms with small genome sizes, such as *Rickettsia prowazekii*, *Borrelia burgdorferi* and *Mycoplasma genitalium*, to nearly 17% in *Synechocystis* PCC6803. In all cases, very ancient horizontal transfer events, such as those disseminating transfer RNA synthetases¹⁷, would not be detected using these methods.

In some species, a substantial proportion of horizontally transferred genes can be attributed to plasmid-, phage- or transposon-related sequences (Fig. 2, yellow bars). As observed in *E. coli*¹, a significant fraction of acquired DNA in *Synechocystis* PCC6803, *Helicobacter pylori* and *Archaeoglobus fulgidus* is physically associated with mobile DNA, which probably mediated the integration of these sequences into the chromosome. Although analyses based on sequence features indicate that large portions of bacterial genomes are attributable to lateral transfer, these methods can underestimate the actual number of transferred genes because sequences acquired from organisms of similar base compositions and codon usage patterns, as well as ancient transfer events, will escape detection.

Comparisons of completely sequenced genomes verify that bacteria have experienced significant amounts of lateral gene transfer,

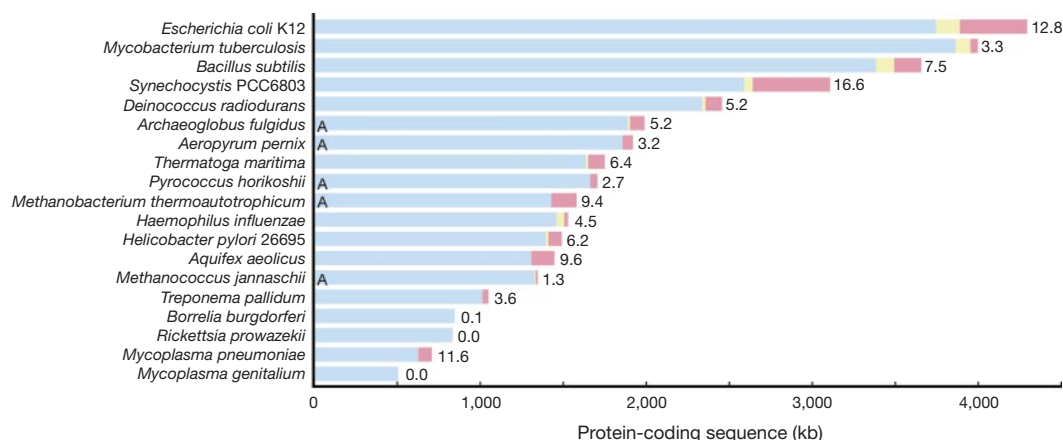


Figure 2 Distribution of horizontally acquired (foreign) DNA in sequenced bacterial genomes. Lengths of bars denote the amount of protein-coding DNA. For each bar, the native DNA is blue; foreign DNA identifiable as mobile elements, including transposons

and bacteriophages, is yellow, and other foreign DNA is red. The percentage of foreign DNA is noted to the right of each bar. 'A' denotes an Archaeal genome.

resulting in chromosomes that are mosaics of ancestral and horizontally acquired sequences. The hyperthermophilic Eubacteria *Aquifex aeolicus* and *Thermotoga maritima* each contain a large number of genes that are most similar in their protein sequences and, in some cases, in their arrangements, to homologues in thermophilic Archaea. Twenty-four per cent of *Thermotoga*'s 1,877 ORFs and about 16% of *Aquifex*'s 1,512 ORFs, display their highest match (as detected by BLAST) to an Archaeal protein, whereas mesophiles, such as *E. coli*, *B. subtilis* and *Synechocystis*, have much lower proportions of genes that are most similar to Archaeal homologues^{18,19}.

Because the detection of lateral transfer by such systematic, gene-by-gene analyses depends upon the occurrence, phylogenetic positions and evolutionary rates of homologues in the currently available databases²⁰, the specific genes, as well as the overall amount of transferred DNA, recognized by this approach could differ from those resolved by examining atypical sequence characteristics. Moreover, the utility of BLAST similarity scores to infer gene ancestry depends upon the number of acknowledged homologues, the evolutionary relationships of organisms represented in the databases and the persistence of genes in a genome; hence, values derived from whole-genome comparisons will be refined as additional genomes are completed²¹. However, both methods reveal the potential for bacterial genomes to incorporate large numbers of unique sequences, often from very divergent organisms.

How and what sequences are acquired

In contrast to the evolution of new traits through the modification of existing sequences, the origin of new abilities through lateral gene transfer has three requirements. First, there needs to be a means for the donor DNA to be delivered into the recipient cell. Second, the acquired sequences must be incorporated into the recipient's genome (or become associated with an autonomous replicating element). And third, the incorporated genes must be expressed in a manner that befits the recipient microorganism. The first two steps are largely indiscriminate with respect to the specific genes or functional properties encoded by the transferred regions, and can occur through three mechanisms: transformation, transduction and conjugation.

(1) Transformation involves the uptake of naked DNA from the environment and has the potential to transmit DNA between very distantly related organisms. Certain bacterial species, such as *Neisseria gonorrhoeae* and *Haemophilus influenzae*, are perpetually competent to accept DNA, whereas others, such as *Bacillus subtilis* and *Streptococcus pneumoniae*, become competent upon reaching a certain physiological stage in their life cycle²².

Effective transformation in *N. gonorrhoeae* and *H. influenzae* requires the presence of specific recognition sequences (5'-GCCGTCTGAA-3' and 5'-AAGTGCAGGT-3', respectively), which are present in their respective genomes at frequencies far greater than those expected at random²³⁻²⁵. Although the presence of specific uptake sequences enhances the transformation efficiency between related species (for example, from *H. parainfluenzae* to *H. influenzae*), many of the naturally competent bacterial species, such as *B. subtilis* and *S. pneumoniae*, do not display sequence preference and are capable of high levels of transformation^{22,26}. But transformation proficiency does not necessarily translate into correspondingly high rates of interspecific gene transfer: although species of *Haemophilus* are naturally competent, the *H. influenzae* Rd genome bears little foreign DNA beyond two large prophages (Fig. 2).

(2) New genetic material can be also introduced into a bacterium by a bacteriophage that has replicated within a donor microorganism and packaged random DNA fragments (generalized transduction) or the DNA adjacent to the phage attachment site (specialized transduction). The amount of DNA that can be transferred in a single event is limited by the size of the phage capsid, but can range upwards of 100 kilobases (kb). Although phages are prevalent in the

environment^{27,28}, the spectrum of microorganisms that can be transduced depends upon receptors recognized by the bacteriophage. Like transformation, transduction does not require donor and recipient cells to be present at the same place, or even the same time. On the other hand, phage-encoded proteins not only mediate the delivery of double-stranded DNA into the recipient cytoplasm, but can also promote the integration of DNA into the chromosome and protect the transferred sequences from degradation by host restriction endonucleases.

(3) Conjugation involves physical contact between donor and recipient cells and can mediate the transfer of genetic material between domains (for example, between bacteria and plants, and between bacteria and yeast)^{29,30}. Typically, DNA is transferred from a donor to a recipient strain by either a self-transmissible or mobilizable plasmid. Conjugation can also mediate the transfer of chromosomal sequences by plasmids that integrate into the chromosome (forming an Hfr strain that can transfer donor DNA to recipient cells), and by conjugative transposons, which encode proteins required for their excision from the donor, formation of a conjugative bridge and transposition into the recipient strain.

Despite the diversity of mechanisms mediating genetic exchange among prokaryotes, the introduction of DNA into a recipient cell's cytoplasm does not ensure successful gene transfer unless the transferred sequences are stably maintained in the recipient microorganism. DNA assimilation into the bacterial genome can exploit one of a number of processes including: (1) persistence as an episome, which requires selection to avoid stochastic loss; (2) homologous recombination, which will serve primarily to reassort variation among closely related taxa and is unlikely to allow introduction of novel traits; (3) integration mediated by bacteriophage integrases or mobile element transposases; and (4) illegitimate incorporation through chance double-strand break repair, as postulated for the integration of mitochondrial sequences into the yeast genome³¹.

Through these mechanisms, virtually any sequence—even those originating in eukaryotes or Archaea—can be transferred to, and between, bacteria. The relatively small sizes of bacterial genomes imply, however, that either the rate of transfer or the maintenance of transferred sequences is very low, or that the maintenance of horizontally transferred sequences is offset by the loss of resident sequences. Bacterial genomes do not contain arbitrary assortments of acquired genes, and the plethora of documented transfer events has provided an abundance of information about the origins and functions of the acquired sequences. Naturally, these studies are biased towards identifying genes that impart both new and consequential functions upon the recipient organism; however, they illustrate both the unprecedented range of mechanisms by which traits are disseminated among bacteria and the impact of lateral gene transfer on bacterial evolution.

Traits introduced through lateral gene transfer

Antibiotic resistance. Antimicrobial resistance genes allow a microorganism to expand its ecological niche, allowing its proliferation in the presence of certain noxious compounds. From this standpoint, it is not surprising that antibiotic resistance genes are associated with highly mobile genetic elements, because the benefit to a microorganism derived from antibiotic resistance is transient, owing to the temporal and spatial heterogeneity of antibiotic-bearing environments. Plasmids are readily mobilizable between taxa and represent the most common method of acquiring antibiotic resistance determinants. As plasmids are rarely integrated into the chromosome, the acquired traits must confer an advantage sufficient to overcome not only inactivation by mutation, but also elimination by segregation.

Transposable elements can also promote the transfer of resistance genes between bacterial genomes. When a resistant determinant, or any gene conferring a selectable phenotype, is flanked by two

insertion sequences, it may be mobilized as a complex transposon; for example, two copies of the insertion sequence IS10 flank a tetracycline resistant determinant and regulatory gene to form transposon Tn10 (ref. 32). Likewise, two IS50 elements flank a three-gene operon that confers resistance to kanamycin, bleomycin and streptomycin forming transposon Tn5 (ref. 33), which can integrate into the chromosomes of phylogenetically diverse bacterial species.

Antibiotic resistance genes can also be propagated by integrons, which are gene expression elements that incorporate promoterless genes, thereby converting them into functional genes. Integrons consist of three elements (Fig. 3): an attachment site where the horizontally acquired sequence is integrated; a gene encoding a site-specific recombinase (that is, integrase); and a promoter that drives expression of the incorporated sequence^{34,35}. A site-specific recombination event between the attachment site and a recombination site, known as the 59-base element, located downstream of a promoterless resistance gene results in the incorporation of the resistance gene and its expression from the integron promoter. Integrase exhibit specificity for their attachment sites but not for the 59-base elements, which allows the stockpiling of many resistance determinants within a given integron. Mobility of integrons demands capture by insertion sequences, transposons or conjugative plasmids³⁶. Aside from antibiotic resistance genes, integrons have recently been implicated in the acquisition of virulence determinants by the cholera-causing bacterium, *Vibrio cholerae*³⁷.

Virulence attributes. Unlike the acquisition of antibiotic resistance, adoption of a pathogenic lifestyle usually involves a fundamental change in a microorganism's ecology. The sporadic phylogenetic distribution of pathogenic organisms has long suggested that bacterial virulence results from the presence (and perhaps acquisition) of genes that are absent from avirulent forms. Evidence for this view has taken several forms, ranging from the discovery of large 'virulence' plasmids in pathogenic *Shigella* and *Yersinia*^{38–41} to the ability to confer pathogenic properties upon laboratory strains of *E. coli* by the experimental introduction of genes from other species^{42,43}.

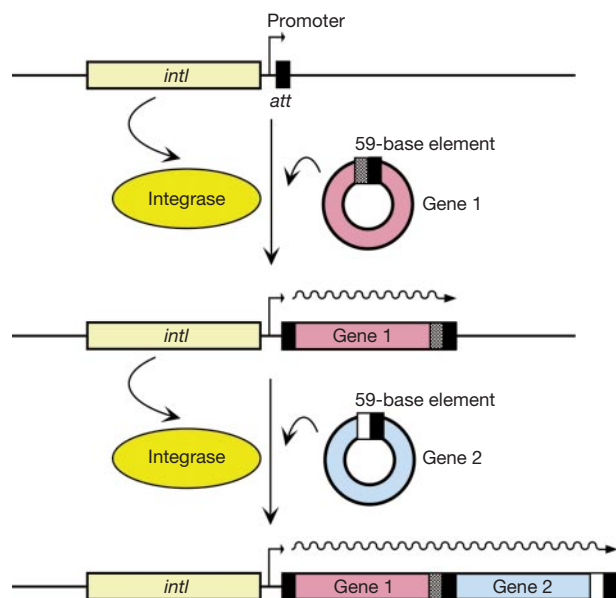


Figure 3 Gene capture and expression by integrons. Integrons consist of an attachment site (*att*), a gene encoding a site-specific recombinase (*intI*) and a promoter that drives expression of the incorporated sequence. Incorporation of a resistance gene (gene 1), and its expression from the integron promoter, result from a site-specific recombination event between the attachment site and a recombination site (known as the 59-base element) located downstream of a promoterless resistance gene. The integrase can recognize different 59-base elements, allowing the stockpiling of many resistance determinants within a given integron.

Recent studies have discovered that horizontally acquired 'pathogenicity islands' are major contributors to the virulent nature of many pathogenic bacteria^{44,45}. These chromosomally encoded regions typically contain large clusters of virulence genes and can, upon incorporation, transform a benign organism into a pathogen. Many pathogenicity islands are situated at tRNA and tRNA-like loci, which appear to be common sites for the integration of foreign sequences. For example, the tRNA^{serC} has repeatedly served as the integration site of pathogenicity islands in enteric bacteria⁴⁶, including the 70-kb PAI-1 of uropathogenic *E. coli*⁴⁷, the 35-kb LEE island of enteropathogenic *E. coli*⁴³, the 24-kb SHI-2 island of *Shigella flexneri*^{48,49} and the 17-kb SPI-3 island of *S. enterica*⁵⁰ (Table 1).

The sequences flanking pathogenicity islands frequently contain short direct repeats reminiscent of those generated upon integration of mobile genetic elements, and ORFs within certain pathogenicity islands display sequence similarity to bacteriophage integrases. Several phages, including ϕ R73 and P4 of *E. coli*, P22 of *S. enterica* and HP1 of *H. influenzae*, insert at or near tRNA genes, suggesting that pathogenicity islands are transferred and acquired through phage-mediated events, or that their incorporation involves the action of conserved integrases^{51,52}. This is clearly the case in *Staphylococcus aureus*, in which a generalized transducing phage promotes the excision, replication and mobilization of a pathogenicity island harbouring the gene for toxic shock toxin⁵³.

Certain bacteriophages encode virulence determinants within their genomes, and lysogenization by such a bacteriophage results in the 'conversion' of a strain to a pathogenic variant. For example, the genes encoding exotoxin A in *Streptococcus pyogenes*⁵⁴, Shiga toxin in enterohaemorrhagic strains of *E. coli*⁵⁵, and the SopE GTP/GDP exchange factor necessary for host cell invasion by *S. enterica*⁵⁶ are carried by converting bacteriophages. In *V. cholerae*, the cholera toxin genes are encoded within the genome of a filamentous bacteriophage (termed CTX ϕ), which uses as its receptor the intestinal colonization factor TcpA encoded within the VP1 pathogenicity island⁵⁷. VP1 constitutes the genome of another bacteriophage (which is distinct from CTX ϕ), however, and TcpA specifies the coat protein of this phage⁵⁸, such that acquisition of cholera toxin and, hence, virulence requires prior lysogenization by the VP1 phage.

Not only have pathogens originated through the acquisition of sequences by lateral gene transfer, but, in some cases, virulence also depends upon the absence of resident genes that diminish pathogenic potential (Fig. 4). An early example of such a virulence suppressor in enteric bacteria is the surface protease OmpT, which is present in nonpathogenic *E. coli* but absent from *Shigella*. The presence of OmpT attenuates virulence by interfering with the expression of the *Shigella* VirG protein, which is required for intercellular spread⁵⁹. In addition to lacking *ompT*, *Shigellae* also have deletions of the region containing the *cadA* gene, which encodes lysine decarboxylase (Fig. 4). When the *cadA* gene from a benign strain of *E. coli* is introduced into *S. flexneri*, the resulting strain can no longer induce the fluid secretion normally associated with infection⁶⁰.

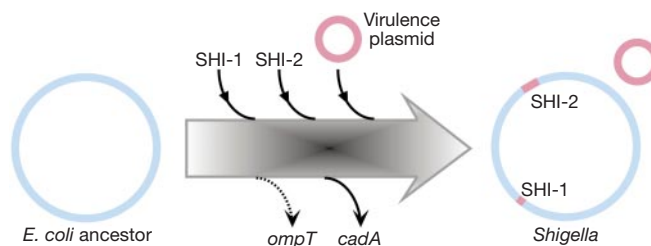


Figure 4 Succession of genetic events contributing to virulence in *Shigella*. Strains of *Shigella* are derived from *E. coli*, and pathogenicity in *Shigella* requires the acquisition of a virulence plasmid and two chromosomally encoded regions as well as the absence of two genes detected in *E. coli*. The deletion of *ompT* is displayed with a dashed line because it is not known whether all *E. coli* ancestors to *Shigella* harbour this phage-borne gene.

Metabolic properties. Lateral gene transfer has played a significant role in moulding bacterial genomes by mobilizing other physiological traits, which have, in effect, allowed recipient organisms to explore new environments. Although not surprising in retrospect, early alignments of the linkage maps of *E. coli* and *S. enterica* showed that many biochemical properties restricted to only one of these species, such as lactose fermentation by *E. coli* or citrate utilization by *S. enterica*, were encoded on chromosomal regions unique to each of these species⁶¹. Subsequent molecular genetic analyses have established that, in many cases, species-specific traits can be attributed either to the acquisition of genes through lateral gene transfer or to the loss of ancestral genes from one lineage. In this way, most ecological innovation in Eubacteria and Archaea is fundamentally different from diversification in multicellular eukaryotes.

The acquisition of new metabolic traits by horizontal transfer emphasizes the importance of natural selection as the arbiter of lateral gene exchange—the duration of acquired sequences will be fleeting if the genes do not contribute a useful or meaningful function upon introduction to the recipient cell. Therefore, we would expect that the successful mobilization of complex metabolic traits requires the physical clustering of genes, such that all necessary genes will be transferred in a single step. As a result, horizontal inheritance will select for gene clusters and for operons, which can be expressed in recipient cells by a host promoter at the site of insertion⁶². In this manner, *E. coli*, by acquiring the *lac* operon, gained the ability to use the milk sugar lactose as a carbon source and to explore a new niche, the mammalian colon, where it established a commensal relationship.

Whole-genome comparisons have uncovered sets of genes that are restricted to organisms that have independently adapted to a common lifestyle, such as Archaeal and Bacterial hyperthermophiles¹⁸ or the intracellular pathogens *Rickettsia* and *Chlamydia*⁶³. The distribution of such sequences in phylogenetically divergent but ecologically similar microorganisms has been ascribed to lateral gene transfer, and it is tempting to suggest that these genes, for which functions are presently unknown, contribute to the metabolic and physiological requirements of these unique environments.

Rate of sequence acquisition

The rate of DNA acquisition by the *E. coli* chromosome was measured indirectly by examining the amelioration of atypical

sequence characteristics (for example, nucleotide composition) towards the equilibrium values displayed by this genome^{1,9}. Such methods allow the estimation of the time of arrival for each segment of foreign DNA detected in the genome and, hence, the rate of successful horizontal genetic transfer. Taking into account the inevitable deletion of genes—bacterial genomes are not growing ever larger in size—horizontal transfer has been estimated to have introduced successfully ~16 kb per million years into the *E. coli* genome.

As evident from Fig. 2, there is broad variation in the amount of acquired DNA among bacterial genomes. Although some organisms may have only a limited capacity to transfer and exchange DNA, bacterial lifestyles can also contribute to the lower rates of gene acquisition in genomes containing small amounts of foreign DNA. For example, the intracellular habitat of *R. prowazekii* and *M. genitalium* probably shields the organism from exposure to potential gene donors and the opportunity to acquire foreign sequences.

The impact of acquired DNA

Lateral gene transfer provides a venue for bacterial diversification by the reassortment of existing capabilities. Yet, while the emergence of new phenotypic properties through lateral gene transfer furnishes several advantages, it also presents several problems to an organism. Newly acquired sequences, especially those conferring traits essential to only a portion of the bacterial life cycle, are most useful when they are appropriately and coordinately regulated with the rest of the genome. In *Salmonella*, the expression of several independently acquired virulence genes is under the control of a single regulatory system—the PhoP/PhoQ two-component system—that was already performing essential functions in the genome before the acquisition of these genes^{50,64–66}. Although the precise manner by which each of these genes is regulated has yet to be resolved, these findings suggest that the physiological capabilities and adaptation of very divergent bacteria rely on a common set of universally distributed regulatory signals.

Despite wide diversity in the structure, organization and contents of bacterial genomes⁶⁷, there is relatively narrow variation in genome sizes. Evidence from experimental studies shows that bacterial genomes are prone towards deleting non-essential DNA, and the small, reduced genomes of host-dependent bacteria, such as *Mycoplasma*, *Chlamydia* and *Rickettsia*, attest to the tendency for

Table 1 Transfer RNA loci targeted by horizontally acquired DNA sequences

tRNA locus	Organism	Horizontally acquired DNA	Trait	Size (kb)
<i>selC</i>	<i>Escherichia coli</i>	Phage ØR73	Phage genome	13
<i>selC</i>	Uropathogenic <i>E. coli</i>	PAI-1	Haemolysin	70
<i>selC</i>	Enteropathogenic <i>E. coli</i>	LEE	Type III secretion system; intimin receptor protein Tir	35
<i>selC</i>	<i>Salmonella enterica</i>	SPI-3	Macrophage survival protein MgtC; Mg ²⁺ transporter MgtB	17
<i>selC</i>	<i>Shigella flexneri</i>	SHI-2	Aerobactin iron transport	24
<i>leuX</i>	Uropathogenic <i>E. coli</i>	PAI-2	Haemolysin; prf fimbriae	190
<i>leuX</i>	<i>E. coli</i>	Phage P4	Phage genome	
<i>leu</i>	<i>Haemophilus influenzae</i>	Phage HPI	Phage genome	
<i>leu</i>	<i>Bacteroides</i> sp.	NBU1		10
<i>pheR</i>	Uropathogenic <i>E. coli</i>	PAI-5	Haemolysin; prs fimbriae; cytotoxic necrotizing factor type 1	110
<i>pheV</i>	Uropathogenic <i>E. coli</i>	PAI-4	Haemolysin; pap fimbriae	170
<i>pheV</i>	<i>E. coli</i>	CTnscr94	Sucrose utilization	100
<i>phe</i>	<i>Mesorhizobium loti</i>	Symbiosis island	Nodulation and nitrogen fixation	500
<i>asnT</i>	<i>Yersinia enterocolitica</i>	HPI	Yersiniabactin	45
<i>asnT</i>	<i>Y. pseudotuberculosis</i>	HPI		
<i>valV</i>	<i>S. enterica</i>	SPI-2	Type III secretion system	40
<i>serT</i>	<i>S. enterica</i>	SPI-5	Inositol phosphate phosphatase	7
<i>serV</i>	<i>Dichelobacter nodosus</i>	Vap region	Putative toxin	12
<i>metV</i>	Uropathogenic <i>E. coli</i>		Haemolysin	50
<i>ssrA</i>	<i>D. nodosus</i>	Vrl region		27
<i>ssrA</i>	<i>Vibrio cholerae</i>	VP1	TcpA colonization factor and receptor for CTX phage	45
<i>thr</i>	<i>S. enterica</i>	Phage P22	Phage genome	44
<i>thr</i>	<i>Listeria ivanovii</i>	Int1 region	Internalins C and D	5
<i>arg</i>	<i>Corynebacterium diphtheriae</i>	Corynebacteriophage	Phage genome; toxin	36
<i>glyV</i>	<i>Pseudomonas</i> sp.	clc element	Chlorocatechol-degradation	105

bacteria to delete the more expendable sequences from their genomes^{68,69}.

Because bacterial genomes can maintain only a finite amount of information against mutation and loss, chromosomal deletions will serve to eliminate genes that fail to provide a meaningful function, that is, the bulk of acquired DNA as well as superfluous ancestral sequences. Hence, bacterial genomes are sampling rather than accumulating sequences, counterbalancing gene acquisition with gene loss^{70,71}. As a result, lateral gene transfer can redefine the ecological niche of a microorganism, which will, in effect, promote bacterial speciation.

A potential result of rampant interspecific recombination is the blurring of species boundaries, and the failure of any one gene to reflect the evolutionary history of the organism as a whole. Yet, as more bacterial genomes are examined, robust consensus phylogenies are being constructed from infrequently transferred genes, which provide the benchmark for gauging the scope and impact of lateral gene transfer. Rather than diminishing the utility of molecular phylogeny, the intermingling of genes and the resulting phylogenetic incongruities document the process of gene-transfer-mediated organismal diversification³. □

- Lawrence, J. G. & Ochman, H. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl Acad. Sci. USA* **95**, 9413–9417 (1998).
- Davies, J. Origins and evolution of antibiotic resistance. *Microbiologia* **12**, 9–16 (1996).
- Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284**, 2124–2129 (1999).
- Swofford, D. L., Olsen, G. J., Wadell, P. J. & Hillis, D. M. in *Molecular Systematics* (eds Hillis, D. M., Moritz, C. & Mable, B. K.) 407–514 (Sinauer Associates, Sunderland, Massachusetts, 1996).
- Sueoka, N. On the genetic basis of variation and heterogeneity in base composition. *Proc. Natl Acad. Sci. USA* **48**, 582–592 (1962).
- Muto, A. & Osawa, S. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl Acad. Sci. USA* **84**, 166–169 (1987).
- Karlin, S., Campbell, A. M. & Mrázek, J. Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.* **32**, 185–225 (1998).
- Groisman, E. A., Saier, M. H. Jr & Ochman, H. Horizontal transfer of a phosphatase gene as evidence for the mosaic structure of the *Salmonella* genome. *EMBO J.* **11**, 1309–1316 (1992).
- Lawrence, J. G. & Ochman, H. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* **44**, 383–397 (1997).
- Lan, R. & Reeves, P. R. Gene transfer is a major factor in bacterial evolution. *Mol. Biol. Evol.* **13**, 47–55 (1996).
- Maynard Smith, J. The detection and measurement of recombination from sequence data. *Genetics* **153**, 1021–1027 (1999).
- Vulic, M., Dionisio, E., Taddei, F. & Radman, M. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in Enterobacteria. *Proc. Natl Acad. Sci. USA* **94**, 9763–9767 (1997).
- Rayssiguier, C., Thaler, D. S. & Radman, M. The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. *Nature* **342**, 396–399 (1989).
- Medigue, C., Rouxel, T., Vigier, P., Henaut, A. & Danchin, A. Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.* **222**, 851–856 (1991).
- Whittam, T. S. & Ake, S. in *Mechanisms of Molecular Evolution* (eds Takahata, N. & Clark, A. G.) 223–246 (Japan Scientific Society Press, Tokyo, 1992).
- Riley, M. & Anilionis, A. Evolution of the bacterial genome. *Annu. Rev. Microbiol.* **32**, 519–560 (1978).
- Wolf, Y. I., Aravind, L., Grishin, N. V. & Koonin, E. V. Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.* **9**, 689–710 (1999).
- Aravind, L., Tatusov, R. L., Wolf, Y. I., Walker, D. R. & Koonin, E. V. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.* **14**, 442–444 (1998).
- Nelson, K. E. et al. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323–329 (1999).
- Logsdon, J. M. Jr & Fuguy, D. M. *Thermotoga* heats up lateral gene transfer. *Curr. Biol.* **9**, R747–R751 (1999).
- Gaasterland, T. Archaeal genomics. *Curr. Opin. Microbiol.* **2**, 542–547 (1999).
- Dubnau, D. DNA uptake in bacteria. *Annu. Rev. Microbiol.* **53**, 217–244 (1999).
- Goodman, S. D. & Socca, J. J. Identification and arrangement of the DNA sequence recognized in specific transformation of *Neisseria gonorrhoeae*. *Proc. Natl Acad. Sci. USA* **85**, 6982–6986 (1988).
- Elkins, C., Thomas, C. E., Seifert, H. S. & Sparling, P. F. Species-specific uptake of DNA by gonococci is mediated by a 10-base-pair sequence. *J. Bacteriol.* **173**, 3911–3913 (1991).
- Smith, H. O., Tomb, J. F., Dougherty, B. A., Fleischmann, R. D. & Venter, J. C. Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. *Science* **269**, 538–540 (1985).
- Davison, J. Genetic exchange between bacteria in the environment. *Plasmid* **42**, 73–91 (1999).
- Jiang, S. C. & Paul, J. H. Gene transfer by transduction in the marine environment. *Appl. Environ. Microbiol.* **64**, 2780–2787 (1998).
- Schickmaier, P. & Schmieger, H. Frequency of generalized transducing phages in natural isolates of the *Salmonella typhimurium* complex. *Appl. Environ. Microbiol.* **61**, 1637–1640 (1995).
- Buchanan-Wollaston, V., Passiatore, J. E. & Canon, F. The *mob* and *oriT* mobilization functions of a bacterial plasmid promote its transfer to plants. *Nature* **328**, 170–175 (1987).
- Heinemann, J. A. & Sprague, G. F. J. Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature* **340**, 205–209 (1989).
- Richetti, M., Fairhead, C. & Dujon, B. Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* **402**, 96–100 (1999).
- Kleckner, N. in *Mobile DNA* (eds Berg, D. E. & Howe, M. M.) 227–268 (American Society for Microbiology, Washington DC, 1989).
- Berg, D. E. in *Mobile DNA* (eds Berg, D. E. & Howe, M. M.) 185–210 (American Society for Microbiology, Washington DC, 1989).
- Hall, R. M. Mobile gene cassettes and integrons: moving antibiotic resistance genes in Gram-negative bacteria. *CIBA Found. Symp.* **207**, 192–205 (1997).
- Rowe-Magnus, D. A. & Mazel, D. Resistance gene capture. *Curr. Opin. Microbiol.* **2**, 483–488 (1999).
- Hall, R. M. & Collis, C. M. Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. *Mol. Microbiol.* **15**, 593–600 (1995).
- Mazel, D., Dychinco, B., Webb, V. A. & Davies, J. A distinctive class of integron in the *Vibrio cholerae* genome. *Science* **280**, 605–608 (1998).
- Portnoy, D. A., Moseley, S. L. & Falkow, S. Characterization of plasmids and plasmid-associated determinants of *Yersinia enterocolitica* pathogenesis. *Infect. Immun.* **51**, 775–782 (1981).
- Maurelli, A. T., Baudry, B., d'Hauteville, H., Hale, T. L. & Sansonetti, P. J. Cloning of plasmid DNA sequences involved in invasion of HeLa cells by *Shigella flexneri*. *Infect. Immun.* **49**, 164–171 (1985).
- Sasakawa, C. et al. Virulence-associated genetic regions comprising 31 kilobases of the 230-kilobase plasmid in *Shigella flexneri* 2a. *J. Bacteriol.* **170**, 2480–2484 (1988).
- Gemski, P., Lazere, J. R., Casey, T. & Wohlhietter, J. A. Presence of virulence-associated plasmid in *Yersinia pseudotuberculosis*. *Infect. Immun.* **28**, 1044–1047 (1980).
- Isberg, R. R. & Falkow, S. A single genetic locus encoded by *Yersinia pseudotuberculosis* permits invasion of cultured animal cells by *Escherichia coli* K-12. *Nature* **317**, 19–25 (1985).
- McDaniel, T. K. & Kaper, J. B. A cloned pathogenicity island from enteropathogenic *Escherichia coli* confers the attaching and effacing phenotype on *E. coli* K-12. *Mol. Microbiol.* **23**, 399–407 (1997).
- Groisman, E. A. & Ochman, H. Pathogenicity islands: bacterial evolution in quantum leaps. *Cell* **87**, 791–794 (1996).
- Hacker, J., Blum-Oehler, G., Muhldorfer, I. & Tschape, H. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol. Microbiol.* **23**, 1089–1097 (1997).
- Ritter, A. et al. tRNA genes and pathogenicity islands: influence on virulence and metabolic properties of uropathogenic *Escherichia coli*. *Mol. Microbiol.* **17**, 109–121 (1995).
- Blum, G. et al. Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an *Escherichia coli* wild-type pathogen. *Infect. Immun.* **62**, 606–614 (1994).
- Moss, J. E., Cardozo, T. J., Zychlinsky, A. & Groisman, E. A. The *selC*-associated SHI-2 pathogenicity island of *Shigella flexneri*. *Mol. Microbiol.* **33**, 74–83 (1999).
- Vokes, S. A., Reeves, S. A., Torres, A. G. & Payne, S. M. The aerobactin iron transport system genes in *Shigella flexneri* are present within a pathogenicity island. *Mol. Microbiol.* **33**, 63–73 (1999).
- Blanc-Potard, A. B. & Groisman, E. A. The *Salmonella selC* locus contains a pathogenicity island mediating intramacrophage survival. *EMBO J.* **16**, 5376–5385 (1997).
- Sun, J., Inouye, M. & Inouye, S. Association of a retroelement with a P4-like cryptic prophage (retrophage Theta713) integrated into the selenocystyl-tRNA gene of *Escherichia coli*. *J. Bacteriol.* **173**, 171–181 (1991).
- Cheetham, B. F. & Katz, M. E. A role for bacteriophages in the evolution and transfer of bacterial virulence determinants. *Mol. Microbiol.* **18**, 201–208 (1995).
- Lindsay, J. A., Ruzin, A., Ross, H. F., Kurepina, N. & Novick, R. P. The gene for toxic shock toxin is carried by a family of mobile pathogenicity islands in *Staphylococcus aureus*. *Mol. Microbiol.* **29**, 527–543 (1998).
- Weeks, C. R. & Ferretti, J. J. The gene for type A streptococcal exotoxin (erythrogenic toxin) is located in bacteriophage T12. *Infect. Immun.* **46**, 531–536 (1984).
- Jackson, M. P., Neill, R. J., O'Brien, A. D., Holmes, R. K. & Newland, J. W. Nucleotide sequence analysis and comparison of the structural genes for Shiga-like toxin-I and Shiga-like toxin II encoded by bacteriophages from *Escherichia coli*. *FEMS Microbiol. Lett.* **44**, 109–114 (1987).
- Mirold, S. Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic *Salmonella typhimurium* strain. *Proc. Natl Acad. Sci. USA* **96**, 9845–9850.
- Waldor, M. K. & Mekalanos, J. J. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* **272**, 1910–1914 (1996).
- Karaolis, D. K., Somara, S., Maneval, D. R. J., Johnson, J. A. & Kaper, J. B. A bacteriophage encoding a pathogenicity island, a type-IV pilus and a phage receptor in cholera bacteria. *Nature* **399**, 675–679 (1999).
- Nakata, N. et al. The absence of a surface protease, OmpT, determines the intercellular spreading ability of *Shigella*: the relationship between the *ompT* and *kcpA* loci. *Mol. Microbiol.* **9**, 459–468 (1993).
- Maurelli, A. T., Fernández, R. E., Bloch, C. A., Rode, C. K. & Fasano, A. “Black holes” and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **95**, 3943–3948 (1998).
- Riley, M. & Sanderson, K. E. in *The Bacterial Chromosome* (eds Riley, M. & Drlica, K.) 85–96 (American Society for Microbiology, Washington DC, 1990).
- Lawrence, J. G. & Roth, J. R. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**, 1843–1860 (1996).
- Wolf, Y. I., Aravind, L. & Koonin, E. V. Rickettsiae and Chlamydia: evidence of horizontal gene transfer and gene exchange. *Trends Genet.* **14**, 442–444 (1998).
- Groisman, E. A. & Ochman, H. How *Salmonella* became a pathogen. *Trends Microbiol.* **5**, 343–349 (1997).
- Groisman, E. A. The ins and outs of virulence gene expression: Mg²⁺ as a regulatory signal. *Bioessays* **20**, 96–101 (1998).
- Deiwick, J., Nikolaus, T., Erdogan, S. & Hensel, M. Environmental regulation of *Salmonella* pathogenicity island 2 gene expression. *Mol. Microbiol.* **31**, 1759–1773 (1999).
- Casjens, S. The diverse and dynamic structure of bacterial genomes. *Annu. Rev. Genet.* **32**, 339–377 (1998).
- Andersson, S. G. E. & Kurland, C. G. Reductive evolution of resident genomes. *Trends Microbiol.* **6**, 263–268 (1998).
- Andersson, J. O. & Andersson, S. G. E. Insights into the evolutionary process of genome degradation. *Curr. Opin. Genet. Dev.* **9**, 664–671 (1999).
- Lawrence, J. G. Gene transfer, speciation, and the evolution of bacterial genomes. *Curr. Opin. Microbiol.* **2**, 519–523 (1999).
- Lawrence, J. G. & Roth, J. R. in *Organization of the Prokaryotic Genome* (ed. Charlebois, R. L.) 263–289 (American Society for Microbiology, Washington, DC, 1999).

Oceanic Cd/P ratio and nutrient utilization in the glacial Southern Ocean

H. Elderfield & R. E. M. Rickaby*

Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ, UK

During glacial periods, low atmospheric carbon dioxide concentration has been associated with increased oceanic carbon uptake, particularly in the southern oceans. The mechanism involved remains unclear. Because ocean productivity is strongly influenced by nutrient levels, palaeo-oceanographic proxies have been applied to investigate nutrient utilization in surface water across glacial transitions. Here we show that present-day cadmium and phosphorus concentrations in the global oceans can be explained by a chemical fractionation during particle formation, whereby uptake of cadmium occurs in preference to uptake of phosphorus. This allows the reconstruction of past surface water phosphate concentrations from the cadmium/calcium ratio of planktonic foraminifera. Results from the Last Glacial Maximum show similar phosphate utilization in the subantarctic to that of today, but much smaller utilization in the polar Southern Ocean, in a model that is consistent with the expansion of glacial sea ice and which can reconcile all proxy records of polar nutrient utilization. By restricting communication between the ocean and atmosphere, sea ice expansion also provides a mechanism for reduced CO₂ release by the Southern Ocean and lower glacial atmospheric CO₂.

Characterization of global ocean nutrient cycling is a central issue for evaluating CO₂ concentrations at glacial times. The excess of nutrients in the modern Southern Oceans results from intense mixing and low nutrient utilization. Therefore, lower glacial polar nutrients, from increased nutrient utilization, increased stratification or both¹, could quantitatively account for the glacial decrease in atmospheric CO₂. Despite much work using multiple proxies of palaeoproductivity, no consensus about the influence of surface waters in the glacial Southern Ocean on atmospheric CO₂ has emerged^{2–8}.

One category of palaeochemical tracers for productivity addresses the flux of organic particles from the surface ocean. These proxies all indicate that there was higher export production in the subantarctic sector of the Southern Ocean (north of the polar front) at glacial times^{6–8}. But there is disagreement concerning whether glacial export productivity was higher, lower or the same as today for the Antarctic sector, south of the polar front^{2,6–8}.

Although export production removes particulate organic carbon from the surface ocean and thus reduces atmospheric CO₂, its overall effect on the atmosphere depends on vertical mixing. If dissolved carbon and associated nutrients are added back to the surface ocean at the same rate as the export flux, the net effect on atmospheric CO₂ will be zero. Therefore, to understand the influence of glacial Southern Ocean surface waters on atmospheric CO₂, we must use proxies of surface water nutrients or nutrient utilization. Here, the debate is greatest. Originally, nutrient levels were estimated from $\delta^{13}\text{C}_{\text{carbonate}}$ of planktonic foraminifera², but disequilibrium effects such as air–sea exchange and carbonate chemistry add complications⁹. Other methods such as Ge/Si in diatoms¹⁰ offered promise but now appear to be unreliable for this purpose¹¹.

At present, four new methods have been applied to this problem, three of which involve fractionation of nutrient isotopes. The formation of marine organic matter fractionates nitrogen^{5,6,12} and carbon^{3,13} isotopes, and the production of biogenic opal fractionates silicon⁴ isotopes. Therefore, records of organic $\delta^{15}\text{N}$, $\delta^{13}\text{C}$ and opal $\delta^{30}\text{Si}$ should all reflect nutrient utilization in the surface ocean, but conflicting results have been produced. For the subantarctic, $\delta^{15}\text{N}$ indicates no glacial–interglacial change^{5,6}, which is consistent with

one interpretation of the latest $\delta^{13}\text{C}_{\text{organic}}$ data³. For the Antarctic, $\delta^{15}\text{N}$ indicates a glacial increase in nitrate utilization^{5,6} whereas $\delta^{13}\text{C}_{\text{organic}}$ indicates that the Antarctic may have been a source of CO₂ at glacial times³, supported by $\delta^{30}\text{Si}$ which indicates less glacial Si utilization⁴ (the Si method is new and no data exist for the subantarctic).

In the fourth method we use the relationship between Cd and P in the ocean, exploited¹⁴ to infer deepwater P from the Cd/Ca of benthic foraminifera, and use planktonic foraminiferal Cd/Ca as a proxy for surface water phosphate utilization¹⁵. Here we show how clarification of the systematics of Cd and P in seawater provides the framework for application of the method. We go on to re-interpret planktonic Cd/Ca records for the Southern Ocean to estimate surface water phosphate at glacial times and discuss its relationship with atmospheric CO₂.

Fractionation between Cd and P in surface waters

The use of Cd in planktonic foraminifera to reconstruct surface water nutrients hinges on an understanding of the relationship between Cd and phosphate in seawater. The current global dataset has been described by two different linear relationships with a ‘kink’ at $\sim 1.3 \mu\text{mol kg}^{-1} \text{P}$ (refs 16–19). We first show that this ‘kink’ is an artefact and that global surface water data can be explained by preferential extraction of Cd relative to P from surface waters, with a constant fractionation factor ($\alpha_{\text{Cd/P}}$).

The dissolved Cd/P ratio decreases from deep to surface waters, a trend that parallels the general profile of nutrients¹⁹ (Fig. 1a) and provides strong evidence for fractionation of Cd from P within the surface ocean. However, the Cd/P ratios of global surface waters show large and previously unexplained variations²⁰, a potential complication in forging a link between planktonic Cd/Ca and surface water phosphate. But our compilation of existing data shows that surface water Cd/P increases from low to high latitudes (Fig. 1b). Given the modern distribution of P in the surface ocean, this indicates a possible link between the Cd/P of surface waters and their P utilization. High-latitude surface waters have high P and Cd/P ratios, like deepwater values, whereas low-latitude waters have low P and low Cd/P. Thus, both the vertical and latitudinal distributions of Cd/P and P can be explained by fractionation during photosynthesis with preferential incorporation of Cd relative to P such that particulate organic matter (POM) is enriched in Cd/P relative to surface waters.

* Present address: Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.

Model approaches

This idea of surface water fractionation is not new²¹ but has not been treated globally before. To test this hypothesis as a first-order approach to the global ocean, we construct some simple models which incorporate surface water fractionation and compare the model results with data for Cd and P in the global ocean, and simulated by a five-box model representation of the modern ocean. We then discuss different oceanic regimes with a more detailed approach.

Rayleigh model. The simplest way to view the fractionation of Cd relative to P is through a closed system approach, as is used for nutrient isotopes^{4,12}. The initial composition of the reservoir reflects the chemistry of the most enriched global deep seawater such that surface water can be derived from this reservoir after fractionation. Fractionation occurs actively through uptake by phytoplankton and/or passively during particle formation^{19,21,22}. Particles (POM) are formed with $[Cd/P]_{POM}/[Cd/P]_{SW} = \alpha_{Cd/P}$ where $\alpha_{Cd/P} > 1$ and reflects the preferential incorporation of Cd. As the initial deep seawater P is progressively utilized during biological uptake, $[Cd/P]_{SW}$ decreases, as predicted by Rayleigh fractionation kinetics and consistent with observations (Fig. 2a). As a result, $[Cd/P]_{POM}$ also decreases, tending towards the initial deepwater (DSW) composition when all the P is utilized. Combining the Rayleigh equations for this system, we obtain

$$[Cd/P]_{SW} = [Cd/P]_{DSW} \times \frac{1}{\alpha_{Cd/P}} \frac{(1 - f^{\alpha_{Cd/P}})}{1 - f} \quad \text{where } f = \frac{[P]_{SW}}{[P]_{DSW}} \quad (1)$$

where a decrease in f is a measure of increasing nutrient utilization

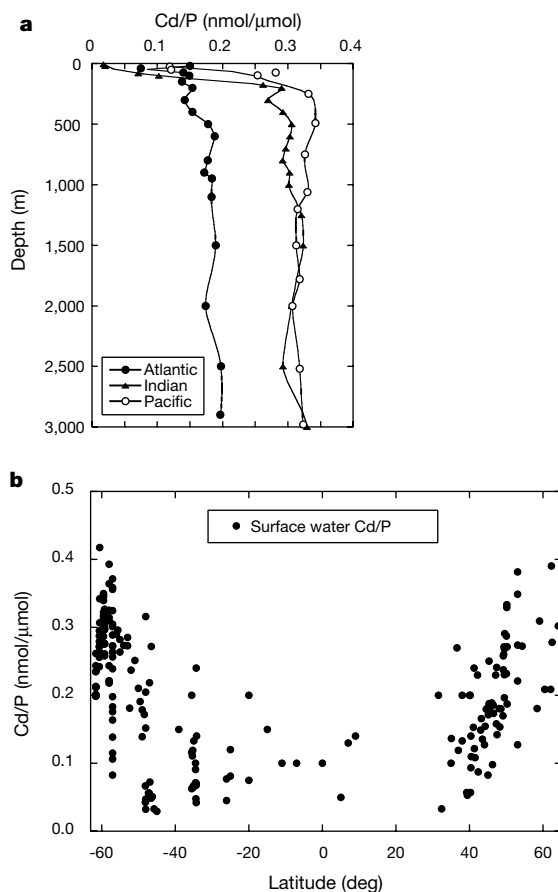


Figure 1 Cd/P ratio in ocean waters. **a**, Depth profiles of Cd/P from the Atlantic⁵⁰ (closed circles), Indian²⁴ (closed triangles) and Pacific oceans²⁶ (open circles). **b**, A compilation of good quality^{16,17,23} global surface seawater Cd/P (nmol/μmol) ratios plotted against latitude.

and removal from seawater by particles. Application of this approach to a compilation of $[Cd/P]_{SW}$ from the Atlantic and Pacific oceans shows that the dataset²³ may be approximated by an $\alpha_{Cd/P}$ of about 2 (Fig. 2a). It is interesting to note the contrast

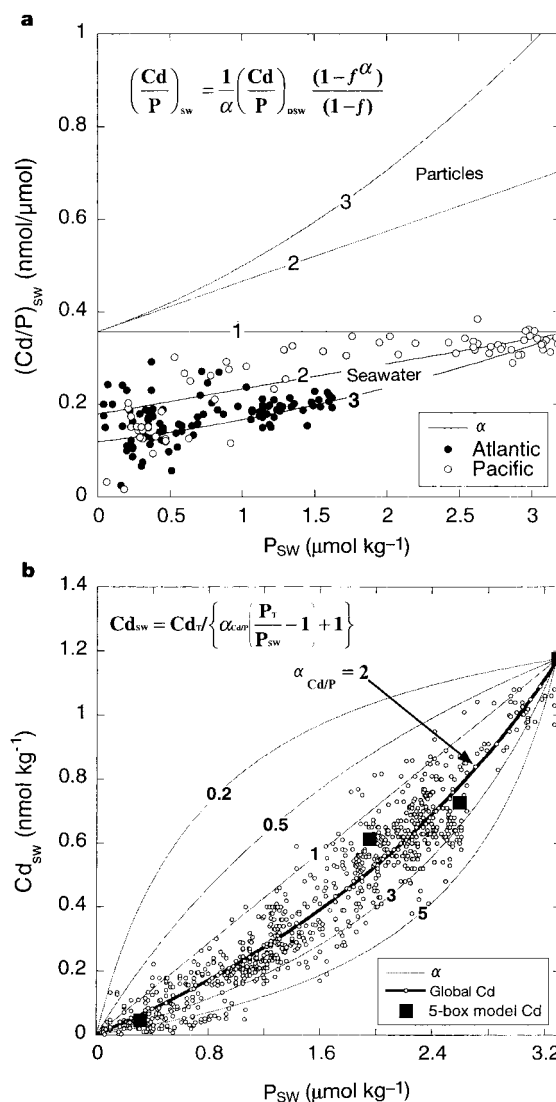


Figure 2 Comparison of model predictions of Cd–P fractionation with oceanic data. **a**, Comparison of $[Cd/P]_{SW}$ plotted against $[P]_{SW}$ for a compilation of $[Cd/P]_{SW}$ ratios from the Atlantic (closed circles) and Pacific (open circles) oceans²³ and the predictions of a closed-system Rayleigh fractionation model (solid lines). The model assumes a globally constant fractionation factor $\alpha_{Cd/P}$ between seawater and particles (POM) and is tested for $\alpha_{Cd/P} = 1-3$. Fractionation is initiated from a reservoir with the geochemical properties of the extreme of the modern global ocean such that $[Cd/P]_{DSW} = 0.36 \text{ nmol/μmol}$ and f is calculated assuming $[P]_{DSW} = 3.3 \text{ μmol kg}^{-1}$. The line of best fit to $[Cd/P]_{SW}$ of the global ocean predicted by Rayleigh fractionation is $\sim \alpha_{Cd/P} = 2$. The POM curves predicted by $\alpha_{Cd/P} = 2$ and 3 are also plotted. When $\alpha_{Cd/P} = 2$, the particulate curve agrees well with the observed range²². **b**, A comparison of seawater Cd and P data (open circles) for the global ocean^{16,17,23} (compiled according to the selection criteria of ref. 16) with the predictions from a two-box mass balance model and a five-box ocean model. The line of best fit to the global data is calculated using equation (2) defined by the mass balance approach and predicts that $\alpha_{Cd/P} = 2$, $[Cd]_T = 1.2 \text{ nmol kg}^{-1}$, $[P]_T = 3.3 \text{ μmol kg}^{-1}$ for the global ocean (thick solid line). The output from each ocean box in our five-box model simulation with $\alpha_{Cd/P} = 2$ is scaled to the Pacific deepwater end-member and plotted (solid squares in order of increasing phosphate: low-latitude surface, subantarctic surface, Antarctic surface and deep). The output from the five-box model agrees well with the global data set. As a comparison, equation (2) using various $\alpha_{Cd/P}$ values with $[Cd]_T$ and $[P]_T$ predicted by the global dataset is plotted as the labelled lines.

between North Atlantic and Pacific deep waters (see below).

Two-box model. Next, we follow the general approach of ref. 10, in which fractionation between Ge and Si in surface waters was described using a two-box model, incorporating mixing of waters between the surface and deep water (at a rate v) and the formation of particles in the surface ocean. P and Cd from the surface ocean are incorporated into particles according to a first-order rate equation (rate constant k) for removal into POM. The particles fall into the

deep ocean and regenerate Cd and P into dissolved form with no further fractionation. The change in the concentration in the surface ocean (shown here for Cd) is described by

$$V_s \frac{d[\text{Cd}]_s}{dt} = v([\text{Cd}]_D - [\text{Cd}]_s) - k_{\text{Cd}}[\text{Cd}]_s V_s$$

where V_s is the volume of the surface ocean and $k_{\text{Cd}}/k_{\text{P}} = \alpha_{\text{Cd/P}}$ At steady state

$$\alpha_{\text{Cd/P}} = \frac{\frac{[\text{Cd}]_T}{[\text{Cd}]_s} - 1}{\frac{[\text{P}]_T}{[\text{P}]_s} - 1} \quad (2)$$

Applying this equation to the global Cd–P relationship (with a floating $[\text{Cd}]_T$ and $[\text{P}]_T$), the model also predicts $\alpha_{\text{Cd/P}} = 2$ with best-fit averaged total Cd and P in the oceans of $[\text{Cd}]_T = 1.2 \text{ nmol kg}^{-1}$ and $[\text{P}]_T = 3.3 \mu\text{mol kg}^{-1}$ (Fig. 2b). The observed ‘kink’ at $1.3 \mu\text{mol kg}^{-1}$ P is simply an artefact of fitting seawater data to two straight lines.

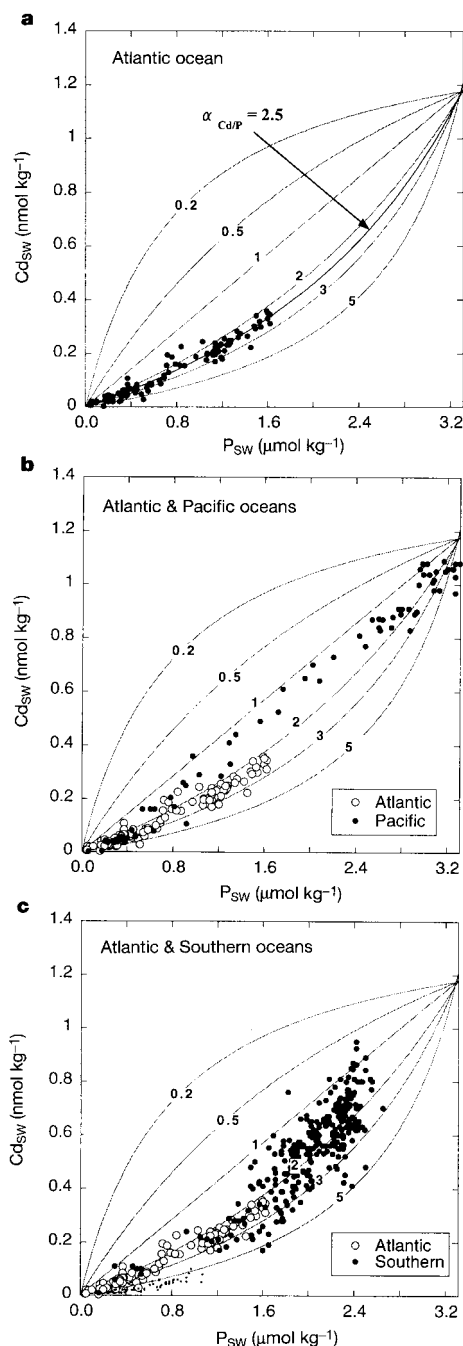


Figure 3 Inter-ocean contrasts in Cd/P compared with fractionation model. Comparison of $\alpha_{\text{Cd/P}}$ for a compilation of surface and deep water Cd and P data from **a**, Atlantic (closed circles); **b**, Atlantic (open circles) and Pacific (closed circles); and **c**, Atlantic (open circles) and Southern ocean (closed circles). The small symbols show the disputed 1992 data set of ref. 25 which has been criticised¹⁶ because P values are higher than for a nearby GEOSECS station and therefore may be in error, but this has been refuted²⁵. As in Fig. 2b, the curves of Cd against P for $\alpha_{\text{Cd/P}} = 0.2$ –5.0 predicted from the mass balance approach using equation (2) are shown as a comparison to the seawater data.

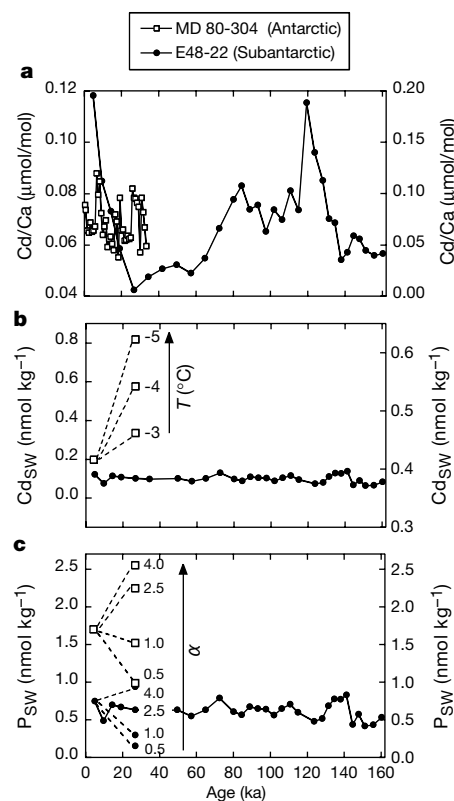


Figure 4 Calculation of Cd_{sw} and P_{sw} from glacial–interglacial records of planktonic foraminiferal Cd/Ca. **a**, Down-core records of Cd/Ca analysed in *Globigerina bulloides* from Subantarctic core E48-22 (39° S, 85° E)¹⁵ (closed circles; left-hand scale), and in *Neogloboquadrina pachyderma* (s) from Antarctic core MD 80-304 (51° S, 67° E)³⁰ (open squares; right-hand scale). A simple linear age model has been constructed from oxygen isotopes for MD 80-304. **b**, Cd_{sw} is calculated by correcting for temperature using the calibration for D_{Cd} (ref. 15), which has now been revised according to the systematics of Cd and P in surface waters to yield: $D_{\text{Cd}} = 0.75 \exp(0.167T)$. SST for E48-22 is derived from Mg/Ca and the modern analogue technique (MAT transfer function) determined on the same core (closed circles)¹⁵. The glacial reduction in SST for MD80-304 is constrained by the $\delta^{18}\text{O}$ record to lie between 3 and 5 °C and we present an LGM–Holocene contrast for the Antarctic for this range (open squares). **c**, P_{sw} concentrations are calculated for the Antarctic (open squares) assuming a temperature difference of –4 °C and the Subantarctic (closed circles) using equation (2). We also show the LGM–Holocene contrast in P_{sw} obtained at each location if glacial $\alpha_{\text{Cd/P}}$ varied from 0.5 to 4.0.

Five-box model. Finally, we have used a five-box ocean–atmosphere model as an analogue of the modern global dataset to simulate Cd and P data for different values of $\alpha_{Cd/P}$. The model²³ is based on that of ref. 1, with a modification to the high-latitude box, which is divided to represent the subantarctic and Antarctic regions of the Southern Ocean, which we discuss later. Again, the best fit was optimized when $\alpha_{Cd/P} = 2$ (Fig. 2b). Note that results for the two high-latitude surface boxes fall at each side of the curve from equation (2) for $\alpha_{Cd/P} = 2$; this simply reflects the definition of a boundary between the surface water boxes which is not represented in the two-box model approach.

Contrasting oceanic regimes

Our approach shows that Cd and P in seawater may be approximated globally by a single value of $\alpha_{Cd/P} = 2$ for the fractionation between Cd and P. However, this is oversimplistic because we have considered surface and deep waters of the ocean together, whereas Cd and P are fractionated in the surface ocean but regenerated in the deep ocean. Fractionation during the formation of POM will deplete surface seawater in Cd relative to P but will tend to enrich deep waters relatively in Cd owing to the regeneration of high Cd/P POM. This is exemplified by considering oceanic regimes separately (Fig. 3).

Surface waters and ‘young’ nutrient-depleted deep waters of the North Atlantic Ocean are accurately represented by our model of fractionation during particle formation, but the curve of best fit is predicted by $\alpha_{Cd/P} = 2.5$ (Fig. 3a), higher than that inferred when both young and older deep waters are considered (Fig. 2). When the ‘old’ and nutrient-rich deepwaters of the Pacific are included they appear to have undergone no fractionation and are best approximated by $\alpha_{Cd/P} = 1.0$ (Fig. 3b). When waters of the Southern Ocean are included, they fall between the two extremes of surface and deep Pacific waters (Fig. 3c).

Thus, the distinctive signatures of these different water masses reflect the combination of Cd/P fractionation in surface waters, regeneration in deep waters and global thermohaline circulation. Surface waters everywhere show Cd/P fractionation with $\alpha_{Cd/P}$ of about 2.5. As POM is regenerated in the deep ocean it increases Cd/P of the deep water. Therefore, the Cd/P signatures of deep waters reflect a mixture of the initial or ‘preformed’ Cd/P and regenerated Cd/P from the decay of particles raining down from the surface ocean.

At the start of the global thermohaline circulation, the deep waters of the Atlantic collect Cd and phosphate, but do not amass enough high Cd/P particles to significantly shift their apparent $\alpha_{Cd/P}$ away from the surface water value. On reaching the high productivity region of the Southern Ocean, high export production increases deepwater Cd/P such that the effective $\alpha_{Cd/P}$ starts to decrease and tends towards 1. This effect is magnified because the particles are forming from high-phosphate surface waters and therefore have higher Cd/P than elsewhere in the global surface ocean (see Fig 2a.). At the end of the thermohaline conveyor in the deep Pacific, the deepwaters have accumulated sufficient Cd/P relative to P that the $\alpha_{Cd/P}$ is restored to 1, the value for the seawater reservoir with no fractionation.

Although this concept of fractionation in surface waters and regeneration in deep waters explains the overall properties of the global ocean, some scatter still remains in the global dataset. Part of this is undoubtedly analytical. The remainder probably results from local factors, or where upwelling of shallower waters occurs^{24,25}. Here, the assumption of a deepwater endmember is not realistic. There may also be regional variations in biomass Cd/P. These appear to be second-order effects and a single $\alpha_{Cd/P}$ of about 2.5 is a good representation of surface waters globally, with no systematic correlation of any major oceanic region with a distinctly different $\alpha_{Cd/P}$ (Figs 2a, 3b).

Given this insight into the systematics of the ocean chemistry of

Cd and P, we can infer dissolved P concentrations directly from foraminiferal Cd/Ca using an $\alpha_{Cd/P}$ of 2.5, provided that the $\alpha_{Cd/P}$ for global surface waters did not change greatly from its present average value. Of course, the effect of the regional deepwater $\alpha_{Cd/P}$ on interpretations of existing benthic foraminiferal Cd/Ca data is very small, as the categorization of deepwater Cd/P ratios based on $1.3 \mu\text{mol kg}^{-1}$ P is comparable to our different deepwater α values.

Mechanism of Cd/P fractionation

Our approach is empirical and says nothing about the mechanisms causing Cd/P fractionation, but simply that it approximates to an average $\alpha_{Cd/P}$ of 2.5 in the modern ocean. This result is consistent with POM having Cd/P above the deep ocean value^{19,21,22} which is a necessary consequence of $\alpha_{Cd/P} > 1$. Measurements of coexisting particles and seawater from Baja California give $\alpha_{Cd/P} = 2.5$ (ref. 26), the same as our estimate. By contrast, laboratory culture studies using diatoms indicate that $\alpha_{Cd/P} < 1$ (ref. 27). However, in these laboratory studies, low utilization of Cd may reflect growth in conditions where zinc is plentiful. Under Zn-limiting conditions, which are typical for the global surface oceans, Cd can substitute for Zn in carbonic anhydrase of siliceous phytoplankton²⁸. Also, there is evidence²⁹ for a form of Cd carbonic anhydrase, which is distinct from other cellular Zn carbonic anhydrase. These data represent the first indication that Cd is required for biological productivity in surface waters and can provide an active mechanism for regulating the Cd/P of surface seawater. Of course, the elemental fractionation of Cd and P may be thought of as analogous to the isotope fractionation systematics of carbon, nitrogen and silicon.

Southern Ocean nutrients

We now apply our understanding of the ocean chemistry of Cd and P to calculate glacial phosphate utilization from records of planktonic foraminiferal Cd/Ca in the Southern Ocean^{15,30–32}. The positions of the cores range from the subtropical frontal zone to Antarctic latitudes, but here we categorize the cores as subantarctic or Antarctic because the same results are found for all cores north and for all cores south of the polar front, respectively.

Planktonic foraminiferal Cd/Ca from the Southern Ocean show a large glacial decrease in the subantarctic^{15,30} but little or no glacial–interglacial contrast in the Antarctic^{31,32} (Fig. 4a). The raw data require correction for the effects of glacial–interglacial changes in sea surface temperature (SST) on D_{Cd} (ref. 15; Fig. 4b) as well as conversion of $[Cd]_{SW}$ to surface water P using $\alpha_{Cd/P}$ (equation (2); Fig. 4c). After these corrections, a very different picture emerges. We see no significant glacial–interglacial change in surface water P in the subantarctic, but an increase in the glacial Antarctic by about $0.6 \mu\text{mol kg}^{-1}$ (Figs 4b, c and 5).

There are two main uncertainties in our calculations of P for the Antarctic. The first is the estimation of glacial SST using diatom transfer functions which can be problematic owing to, for example, selective dissolution leading to biases in the record³³. We therefore show Antarctic P concentrations obtained for different glacial–interglacial ΔSST . Our calculated glacial–interglacial increase in P can be eliminated only if Antarctic ΔSST was zero (Fig. 5). There is convincing evidence from well calibrated transfer functions that glacial SST south of the polar front was $\sim 4^\circ\text{C}$ cooler than at present^{33–35}. Although the absolute values of estimated SST from different functions may differ slightly, there is little impact on D_{Cd} at this lower end of the temperature scale¹⁵. Further support for lower glacial Antarctic SST comes from the Vostok ice core. The air temperature above Antarctica decreased by approximately 6°C at glacial times³⁶ and we can consider SST at the Antarctic core sites as falling on a mixing line between Vostok air temperature and subantarctic SST values further north. This indicates that Antarctic SST may have decreased by about $4\text{--}5^\circ\text{C}$, equivalent to ΔP of about $+0.6 \mu\text{mol kg}^{-1}$.

The second uncertainty is that we have assumed that $\alpha_{\text{Cd/P}}$ is uniform globally and has remained constant. Because we see no regional trend in the modern surface ocean we consider it unlikely that major local changes occurred on a glacial–interglacial timescale. It is possible that the global average may have changed because of shifts in the factors controlling $\alpha_{\text{Cd/P}}$. However, the inferred glacial–interglacial increase in P in the Antarctic can only be eliminated if glacial $\alpha_{\text{Cd/P}} \approx 1.0$ (Fig. 5). A shift in $\alpha_{\text{Cd/P}}$ to about 0.4 is required for our results to be compatible with glacial $\delta^{15}\text{N}$ data south of the polar front. This would require glacial particles to discriminate against Cd such that POM is depleted in Cd/P relative to seawater. Such extreme values of $\alpha_{\text{Cd/P}}$ are not typical of anywhere in the modern surface ocean (Fig. 2a). Benthic Cd/Ca data³⁷ would also require re-interpretation: the inferred P of glacial North Atlantic intermediate (0.7 $\mu\text{mol kg}^{-1}$) and deep (1.5 $\mu\text{mol kg}^{-1}$) waters would change to 0.2 and 0.5 $\mu\text{mol kg}^{-1}$, respectively, both lower than modern values, and create inconsistencies with P reconstructions from benthic $\delta^{13}\text{C}$ (ref. 37). Therefore, our assumption of a glacial $\alpha_{\text{Cd/P}} > 1$ seems reasonable.

Comparison of proxies

Our Cd/Ca results for the subantarctic are consistent with $\delta^{15}\text{N}$ in showing little glacial–interglacial change in nutrient utilization^{5,6,12}. Because of the dependence of $\delta^{13}\text{C}_{\text{org}}$ on diatom growth rate, data³ either show that subantarctic surface waters were in equilibrium with atmospheric pCO_2 (that is, no CO_2 sink, in agreement with Cd/Ca and $\delta^{15}\text{N}$) or reflect a $\sim 20 \mu\text{mol kg}^{-1}$ drawdown in TCO_2 (equivalent to a reduction in P of about 1.2 $\mu\text{mol kg}^{-1}$ assuming modern C/P ratios, completely at odds with Cd/Ca and $\delta^{15}\text{N}$). Therefore, we can say that, with present knowledge, the nutrient tracers are consistent for the subantarctic in showing little or no glacial–interglacial change.

For the glacial Antarctic, our results agree with $\delta^{30}\text{Si}$ (ref. 4), and also $\delta^{13}\text{C}_{\text{org}}$ of diatoms^{3,13} (again, $\delta^{13}\text{C}_{\text{org}}$ is sensitive to growth rate

assumptions), as well as foraminiferal $\delta^{13}\text{C}$ (ref. 13) and opal accumulation rates², but are at odds with the results of $\delta^{15}\text{N}$ (refs 5, 6). Estimates of change in nutrient utilization efficiency ($1 - f$) from glacial to Holocene suggest an increase from about 0–10% to 35% for P (from Cd/Ca), an increase from 40% to 90% for Si (ref. 4) but a decrease from 70% to 30% for nitrate^{5,6} (Fig. 6). Because Cd/Ca follows $\delta^{30}\text{Si}$ and not $\delta^{15}\text{N}$, this indicates that Cd must be tracing P more faithfully than nitrate (there is no evidence of anomalously low Cd relative to global Cd/P in the Arabian Sea denitrification zone^{19,24}).

How, then, can we reconcile $\delta^{15}\text{N}$ with the increasingly consistent picture of high glacial nutrient concentrations in the surface waters of the glacial Southern Ocean? Apart from questioning the interpretation of $\delta^{15}\text{N}$ (for example, the influence of changing rates of N_2 fixation³⁸ versus denitrification³⁹, or recycled versus new production⁴⁰), our results require the glacial N/P to fall to one-third of the modern value. Iron fertilization^{41,42} seems unlikely to be responsible, as Fe-driven N_2 fixation should drive the system towards P limitation⁴³. Alternatively, a shift in the phytoplankton community from predominantly diatoms to *Phaeocystis antarctica* could reduce the N/P ratio by at least a factor of 2 (ref. 44), a change that requires increased glacial upwelling. Instead, we propose that an extension of sea ice at glacial times has the potential to account for all the nutrient proxies and additionally provides a mechanism for the coexistence of higher surface nutrients in the glacial ocean with lower atmospheric CO_2 .

Sea-ice cover in the Southern Ocean at glacial times may have been as much as double the surface area of modern winter ice⁴⁵. If the modern winter surface water P maximum south of the polar front reflects a cessation of productivity because of ice cover⁴⁶ (Fig. 5), then expansion of sea ice at glacial times and diminished productivity can also account for the glacial surface water P inferred from planktonic Cd/Ca. Significant enrichments in ^{15}N have been measured in organic matter that is in or closely associated with sea

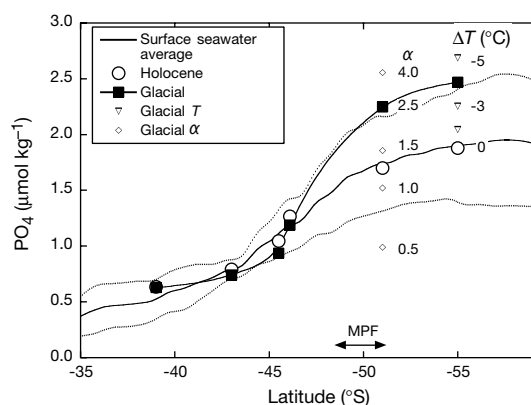


Figure 5 Modern versus LGM reconstruction of Southern Ocean surface water phosphate. A comparison of latitudinal surface water phosphate concentrations for the Subantarctic and Antarctic during the Holocene and the Last Glacial Maximum (LGM) derived from cores E48-22, E49-19, E49-18 (ref. 15), MD 88-769 (ref. 32), MD 80-304 (ref. 30) and All107-22GGC (ref. 31). The position of the modern Polar front (MPF) is marked. Surface water phosphate data are derived from planktonic foraminiferal Cd/Ca for the Holocene (open circles) and the LGM (closed squares). The Cd/Ca records are corrected for temperature using the sequence of calculations outlined for Fig. 3. LGM SST for the subantarctic records is derived from both Mg/Ca and the MAT transfer function. LGM SST for the Antarctic was estimated to be 4 °C cooler than the Holocene from diatom transfer functions. Given that there are some uncertainties in this estimate, we show results at 55°S for a range of LGM–interglacial SST differences (down open triangles). We also show the results obtained if glacial $\alpha_{\text{Cd/P}}$ varied from 0.5 to 4.0 (open diamonds) for 51°S, assuming $\Delta T = -4$ °C. As a means of comparison to the glacial latitudinal profile (thick solid line), the modern seasonal range (dotted line) and annual average surface water P (thick solid line) are plotted⁴⁶.

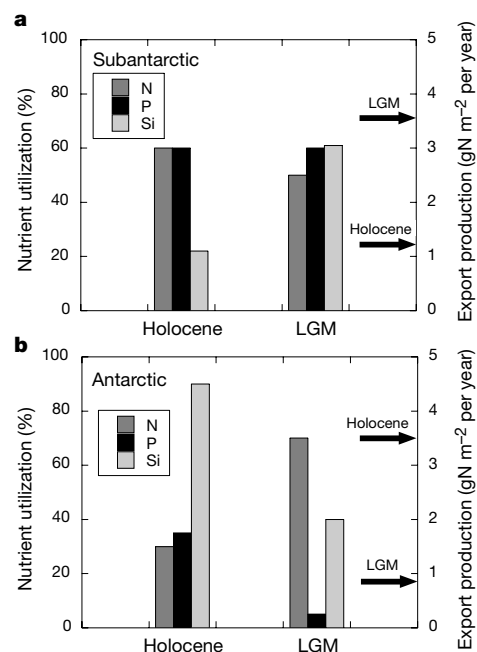


Figure 6 Holocene/LGM contrasts in nutrient utilization. Bar chart showing the per cent surface water utilization of nitrate inferred from $\delta^{15}\text{N}$ (dark grey), phosphate from planktonic foraminiferal Cd/Ca (black), and silicate from $\delta^{30}\text{Si}$ (Antarctic) and roughly estimated from opal accumulation rates (subantarctic) (light grey) for the Holocene and LGM for **a**, the subantarctic and **b**, the Antarctic sectors of the Southern Ocean. Export productivity inferred from particle proxies is also shown for the Holocene and LGM in the subantarctic and Antarctic regions as the horizontal black arrows (note that there is disagreement as to the LGM value for the Antarctic; see text).

ice and consistent with higher recycled production⁴⁰. Thus, the expansion of glacial sea ice can explain the anomalous $\delta^{15}\text{N}$. Finally, and more crucially, increased ice cover acts as a barrier to air–sea exchange, preventing loss of CO_2 from the surface ocean. This process, previously proposed to explain the benthic $\delta^{13}\text{C}$ –Cd/Ca discordancy⁴⁷, has been simulated in box models^{48,49} which show that inhibiting communication between the ocean and atmosphere in the polar regions can account for much of the observed glacial lowering of CO_2 . Therefore, increased ice cover of the glacial Antarctic ocean can unify all strands of evidence. Sea ice hindered productivity in the Southern Ocean, leading to high surface water nutrients with concomitant enrichment in $\delta^{15}\text{N}$ of organic matter, but, ironically, trapped the excess carbon in the deep ocean, thus lowering atmospheric CO_2 at glacial times. □

Received 7 April 1999; accepted 21 February 2000.

1. Toggweiler, J. R. & Sarmiento, J. In *The Carbon Cycle and Atmospheric CO_2 : Natural Variations Archaean to Present* Geophys. Monogr. Ser. Vol. 32 (eds Sundquist, E. T. & Broecker, W. S.) 163–184 (AGU, Washington DC, 1985).
2. Mortlock, R. A. *et al.* Evidence for lower productivity in the Antarctic Ocean during the last glaciation. *Nature* **351**, 220–223 (1991).
3. Rosenthal, Y., Dahan, M. & Shemesh, A. Southern Ocean contributions to glacial–interglacial changes of atmospheric pCO_2 : an assessment of carbon isotope records in diatoms. *Paleoceanography* **15**, 65–75 (2000).
4. De La Rocha, C. L., Brzezinski, M. A., DeNiro, M. J. & Shemesh, A. Silicon-isotope composition of diatoms as an indicator of past oceanic change. *Nature* **395**, 680–683 (1998).
5. Sigman, D., Altabet, M. A., Francois, R., McCorkle, D. C. & Gaillard, J.-F. The isotopic composition of diatom-bound nitrogen in Southern Ocean sediments. *Paleoceanography* **14**, 118–134 (1999).
6. Francois, R. *et al.* Contribution of Southern Ocean surface–water stratification to low atmospheric CO_2 concentrations during the last glacial period. *Nature* **389**, 929–935 (1997).
7. Kumar, N. *et al.* Increased biological productivity and export production in the glacial Southern Ocean. *Nature* **378**, 675–680 (1995).
8. Anderson, R. F. *et al.* Late Quaternary changes in productivity of the Southern Ocean. *J. Mar. Syst.* **17**, 497–514 (1998).
9. Kohfeld, K. E., Anderson, R. F. & Lynch-Stieglitz, J. Carbon isotopic disequilibrium in polar planktonic foraminifera and its impact on modern and last glacial maximum reconstructions. *Paleoceanography* **15**, 53–64 (2000).
10. Murman, R. J. & Stallard, R. F. Germanium/silicon fractionation during biogenic opal formation. *Paleoceanography* **3**, 461–471 (1988).
11. Bareille, G., Labracherie, M., Mortlock, R. A., Maier-Reimer, E. & Froelich, P. N. A test of $(\text{Ge/Si})_{\text{opal}}$ as a palaeorecorder of $(\text{Ge/Si})_{\text{seawater}}$. *Geology* **26**, 179–182 (1998).
12. Francois, R., Bacon, M. P., Altabet, M. A. & Labeyrie, L. D. Glacial/interglacial changes in sediment rain rate in the SW Indian sector of Subantarctic waters as recorded by ^{230}Th , ^{231}Pa , U, and $\delta^{15}\text{N}$. *Paleoceanography* **8**, 611–629 (1993).
13. Shemesh, A., Macko, S. A., Charles, C. D. & Rau, G. H. Isotopic evidence for reduced productivity in the glacial Southern Ocean. *Science* **262**, 407–410 (1993).
14. Boyle, E. A. Cadmium: Chemical tracer of deepwater paleoceanography. *Paleoceanography* **3**, 471–489 (1988).
15. Rickaby, R. E. M. & Elderfield, H. Planktonic foraminiferal Cd/Ca: Paleonutrients or paleotemperature. *Paleoceanography* **14**, 293–303 (1999).
16. de Baar, H. J. W., Saager, P. M., Nolting, R. F. & van der Meer, J. Cadmium versus phosphate in the world ocean. *Mar. Chem.* **46**, 261–281 (1994).
17. Loscher, B. M., van der Meer, J., de Baar, H. J. W., Saager, P. M. & de Jong, J. T. M. The global Cd/P relationship in deep ocean waters and the need for accuracy. *Mar. Chem.* **59**, 87–93 (1997).
18. Yeats, P. A. An isopycnal analysis of cadmium distributions in the Atlantic Ocean. *Mar. Chem.* **61**, 15–23 (1998).
19. Saager, P. M. On the relationship between trace metals and nutrients in sea water. Thesis, Free Univ. (1994).
20. Rutgers van der Loeff, M., Helmers, E. & Kattner, G. Continuous transects of cadmium, copper and aluminium in surface waters of the Atlantic Ocean, 50°N to 50°S : Correspondence and contrast with nutrient-like behaviour. *Geochim. Cosmochim. Acta* **50**, 47–61 (1997).
21. Boyle, E. A., Huested, S. S. & Jones, S. P. On the distribution of copper, nickel and cadmium in the surface waters of the North Atlantic and North Pacific Ocean. *J. Geophys. Res.* **86**, 8048–8066 (1981).
22. Sherrell, R. M. The trace metal geochemistry of suspended oceanic particulate matter. Thesis, MIT (1989).
23. Rickaby, R. E. M. Planktonic foraminiferal Cd/Ca: a new perspective on South Ocean palaeoproductivity. Thesis, Cambridge Univ. (1999).
24. Saager, P. M., de Baar, H. J. W. & Howland, R. J. Cd, Zn, Ni and Cu in the Indian Ocean. *Deep-Sea Res.* **39**, 9–35 (1992).
25. Frew, R. D. & Hunter, K. A. Cadmium–phosphorus cycling at the subtropical convergence south of New Zealand. *Mar. Chem.* **51**, 223–237 (1995).
26. Bruland, K. W., Knauer, G. A. & Martin, J. A. Cadmium in north east Pacific waters. *Limnol. Oceanogr.* **23**, 618–625 (1980).
27. Abe, K. & Matsunaga, K. Mechanism controlling Cd and PO_4 concentrations in Funka Bay, Japan. *Mar. Chem.* **23**, 145–152 (1988).
28. Price, N. M. & Morel, F. M. M. Cadmium and cobalt substitution for zinc in a marine diatom. *Nature* **344**, 658–660 (1990).
29. Cullen, J. T., Lane, T. W., Morel, F. M. M. & Sherrell, R. M. Modulation of cadmium utilization by phytoplankton by seawater concentration. *Nature* **402**, 165–167.
30. Boyle, E. A. The role of vertical chemical fractionation in controlling Late Quaternary atmospheric carbon dioxide. *J. Geophys. Res.* **93**, 15701–15714 (1988).
31. Keigwin, L. D. & Boyle, E. A. Late Quaternary paleochemistry of high-latitude surface waters. *Paleogeogr. Paleoclimatol. Paleocool.* **73**, 85–106 (1989).
32. Rosenthal, Y., Boyle, E. A. & Labeyrie, L. Last glacial maximum paleochemistry and deepwater circulation in the Southern Ocean: Evidence from foraminiferal cadmium. *Paleoceanography* **12**, 787–797 (1997).
33. Pichon, J.-J. *et al.* Surface water temperature changes in the high latitudes of the Southern hemisphere over the last glacial–interglacial cycle. *Paleoceanography* **7**, 289–318 (1992).
34. Luz, B. Late Pleistocene paleoclimates of the South Pacific based on statistical analysis of planktonic foraminifera. *Paleogeogr. Paleoclimatol. Paleocool.* **22**, 61–78 (1977).
35. Zielinski, U., Gersonde, R., Sieger, R. & Utterer, D. Quaternary surface water temperature estimations: Calibration of a diatom transfer function for the Southern Ocean. *Paleoceanography* **13**, 365–383 (1998).
36. Jouzel, J. *et al.* Extending the Vostok ice-core record of paleoclimate to the penultimate glacial period. *Nature* **364**, 411–416 (1993).
37. Boyle, E. A. Oceanic chemical distributions during the stage 2 glacial maximum: Cadmium and $\delta^{13}\text{C}$ evidence compared. *Annu. Rev. Earth Planet. Sci.* **20**, 245–287 (1992).
38. Broecker, W. S. & Henderson, G. M. The sequence of events surrounding Termination II and their implications for the cause of glacial–interglacial CO_2 changes. *Paleoceanography* **13**, 352–364 (1998).
39. Ganeshram, R. S., Pedersen, T. F., Calvert, S. E. & Murray, J. W. Large changes in oceanic nutrient inventories from glacial–interglacial periods. *Nature* **376**, 755–757 (1995).
40. Rau, G. H., Sullivan, C. W. & Gordon, L. I. $\delta^{13}\text{C}$ and $\delta^{15}\text{N}$ variations in Weddell Sea particulate organic matter. *Mar. Chem.* **35**, 355–369 (1991).
41. Hutchins, D. A. & Bruland, K. W. Iron-limited diatom growth and Si:N uptake ratios in a coastal upwelling regime. *Nature* **393**, 561–564 (1998).
42. Takeda, S. Influence of iron availability on nutrient consumption ratio of diatoms in oceanic waters. *Nature* **393**, 774–777 (1998).
43. Falkowski, P. G. Evolution of the nitrogen cycle and its influence on the biological pump in the ocean. *Nature* **342**, 637–642 (1997).
44. Arrigo, K. R. *et al.* Phytoplankton community structure and the drawdown of nutrients and CO_2 in the Southern Ocean. *Science* **283**, 365–367 (1999).
45. Crosta, X., Pichon, J.-J. & Burckle, L. H. Application of modern analog technique to marine Antarctic diatoms: Reconstruction of maximum sea-ice extent at the Last Glacial Maximum. *Paleoceanography* **13**, 284–297 (1998).
46. Levitus, S., Burgett, R. & Boyer, T. *World Ocean Atlas 1994* Vol. 3 Nutrients (Natl Oceanic Atmos. Admin. Atlas Natl Environ. Satellite Data Inf. Serf. 3, US Dept of Comm., Washington DC, 1994).
47. Broecker, W. S. An oceanographic explanation for the apparent carbon isotope–cadmium discordancy in the glacial Antarctic. *Paleoceanography* **8**, 137–139 (1993).
48. Stephens, B. & Keeling, R. F. The influence of Antarctic sea ice on glacial–interglacial CO_2 variations. *Eos* **80**, F79 (1999).
49. Toggweiler, J. R. Variation of atmospheric CO_2 by ventilation of the ocean's deepest water. *Paleoceanography* **14**, 571–588 (1999).
50. Martin, J. H., Fitzwater, S. E., Gordon, R. M., Hunter, C. N. & Tanner, S. J. Iron, primary production and carbon–nitrogen flux studies during the JGOFS North Atlantic Bloom Experiment. *Deep-Sea Res.* **40**, 115–134 (1993).

Acknowledgements

We thank Y. Rosenthal, R. Sherrell and J. Cullen for preprints of their work and N. Shackleton for helpful discussions. This work was supported by NERC.

Correspondence and requests for materials should be addressed to H.E. (e-mail: he101@esc.cam.ac.uk).

The DNA sequence of human chromosome 21

The chromosome 21 mapping and sequencing consortium

M. Hattori*, A. Fujiyama*, T. D. Taylor*, H. Watanabe*, T. Yada*, H.-S. Park*, A. Toyoda*, K. Ishii*, Y. Totoki*, D.-K. Choi*, E. Soeda†, M. Ohki‡, T. Takagi§, Y. Sakaki*§§; S. Taudien||, K. Blechschmidt||, A. Polley||, U. Menzel||, J. Delabar§, K. Kumpf||, R. Lehmann||, D. Patterson#, K. Reichwald||, A. Rump||, M. Schillhabel||, A. Schudy||, W. Zimmermann||, A. Rosenthal||; J. Kudoh*, K. Shibuya*, K. Kawasaki*, S. Asakawa*, A. Shintani*, T. Sasaki*, K. Nagamine*, S. Mitsuyama*, S. E. Antonarakis**, S. Minoshima*, N. Shimizu*, G. Nordsiek††, K. Hornischer††, P. Brandt††, M. Scharfe††, O. Schön††, A. Desario‡‡, J. Reichelt††, G. Kauer††, H. Blöcker††; J. Ramser§§, A. Beck§§, S. Klages§§, S. Hennig§§, L. Riesselmann§§, E. Dagand§§, T. Haaf§§, S. Wehrmeyer§§, K. Borzym§§, K. Gardiner#, D. Nizetic|||, F. Francis§§, H. Lehrach§§, R. Reinhardt§§ & M.-L. Yaspo§§

Consortium Institutions:

* RIKEN, Genomic Sciences Center, Sagamihara 228-8555, Japan

|| Institut für Molekulare Biotechnologie, Genomanalyse, D-07745 Jena, Germany

§ Department of Molecular Biology, Keio University School of Medicine, Tokyo 160-8582, Japan

†† GBF (German Research Centre for Biotechnology), Genome Analysis, D-38124 Braunschweig, Germany

§§ Max-Planck-Institut für Molekulare Genetik, D-14195 Berlin-Dahlem, Germany

Collaborating Institutions:

† RIKEN, Life Science Tsukuba Research Center, Tsukuba 305-0074, Japan

‡ Cancer Genomics Division, National Cancer Center Research Institute, Tokyo 104-0045, Japan

§ Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan

¶ UMR 8602 CNRS, UFR Necker Enfants-Malades, Paris 75730, France

Eleanor Roosevelt Institute, Denver, Colorado 80206, USA

** Medical Genetics Division, University of Geneva Medical School, Geneva 1211, Switzerland

‡‡ CNRS UPR 1142, Institut de Biologie, Montpellier, 34060, France

||| School of Pharmacy, University of London, London WC1N 1AX, UK

Chromosome 21 is the smallest human autosome. An extra copy of chromosome 21 causes Down syndrome, the most frequent genetic cause of significant mental retardation, which affects up to 1 in 700 live births. Several anonymous loci for monogenic disorders and predispositions for common complex disorders have also been mapped to this chromosome, and loss of heterozygosity has been observed in regions associated with solid tumours. Here we report the sequence and gene catalogue of the long arm of chromosome 21. We have sequenced 33,546,361 base pairs (bp) of DNA with very high accuracy, the largest contig being 25,491,867 bp. Only three small clone gaps and seven sequencing gaps remain, comprising about 100 kilobases. Thus, we achieved 99.7% coverage of 21q. We also sequenced 281,116 bp from the short arm. The structural features identified include duplications that are probably involved in chromosomal abnormalities and repeat structures in the telomeric and pericentromeric regions. Analysis of the chromosome revealed 127 known genes, 98 predicted genes and 59 pseudogenes.

Chromosome 21 represents around 1–1.5% of the human genome. Since the discovery in 1959 that Down syndrome occurs when there are three copies of chromosome 21 (ref. 1), about twenty disease loci have been mapped to its long arm, and the chromosome's structure and gene content have been intensively studied. Consequently, chromosome 21 was the first autosome for which a dense linkage map², yeast artificial chromosome (YAC) physical maps^{3–6} and a *NotI* restriction map⁷ were developed. The size of the long arm of the chromosome (21q) was estimated to be around 38 megabases (Mb), based on pulsed-field gel electrophoresis (PFGE) studies using *NotI* restriction fragments⁷. By 1995, when the sequencing effort was initiated, around 60 messenger RNAs specific to chromosome 21 had been characterized. Here we report and discuss the sequence and gene catalogue of the long arm of chromosome 21.

Chromosome geography

Mapping. We converted the euchromatic part of chromosome 21 into a minimum tiling path of 518 large-insert bacterial clones. This collection comprises 192 bacterial artificial chromosomes (BACs), 111 P1 artificial chromosomes (PACs), 101 P1, 81 cosmids, 33 fosmids and 5 polymerase chain reaction (PCR) products (Fig. 1). We used clones originating from four whole-genome libraries and nine chromosome-21-specific libraries. The latter were particularly useful for mapping the centromeric and telomeric repeat-contain-

ing regions and sequences showing homology with other human chromosomes.

We used two strategies to construct the sequence-ready map of chromosome 21. In the first, we isolated clones from arrayed genomic libraries by large-scale non-isotopic hybridization⁸. We built primary contigs from hybridization data assembled by simulated annealing, and refined clone overlaps by restriction digest fingerprinting. Contigs were anchored onto PFGE maps of *NotI* restriction fragments and ordered using known sequence tag site (STS) framework markers. We used metaphase fluorescent *in situ* hybridization (FISH) to check the locations of more than 250 clones. The integrity of the contigs was confirmed by FISH, and gaps were sized by a combination of fibre FISH and interphase nuclei mapping. Gaps were filled by multipoint clone walking. In the second strategy, we isolated seed clones using selected STS markers and then either end-sequenced or partially sequenced them at fivefold redundancy. Seed clones were extended in both directions with new genomic clones, which were identified either by PCR using amplimers derived from parental clone ends or by sequence searches of the BAC end sequence database (<http://www.tigr.org>). Nascent contigs were confirmed by sequence comparison.

The final map is shown in Fig. 1. It comprises 518 bacterial clones forming four large contigs. Three small clone gaps remain despite screening of all available libraries. The estimated sizes of

these gaps are 40, 30 and 30 kilobases (kb), respectively, as indicated by fibre FISH (see supporting data set, last section (<http://chr21.r2-berlin.mpg.de>)).

Sequencing. We used two sequencing strategies. In the first, large-insert clones were shotgun cloned into M13 or plasmid vectors. DNA of subclones was prepared or amplified, and then sequenced using dye terminator and dye primer chemistry. On average, clones were sequenced at 8–10-fold redundancy. In the second approach, we sequenced large-insert clones using a nested deletion method⁹. The redundancy of the nested deletion method was about fourfold. Gaps were closed by a combination of nested deletions, long reads, reverse reads, sequence walks on shotgun clones and large insert clones using custom primers. Some gaps were also closed by sequencing PCR products.

The total length of the sequenced parts of the long arm of chromosome 21 is 33,546,361 bp. The sequence extends from a 25-kb stretch of α -satellite repeats near the centromere to the telomeric repeat array. Seven sequencing gaps remain, totalling less than 3 kb. The largest contig spans 25.5 Mb on 21q. The total length of 21q, including the three clone gaps, is about 33.65 Mb. Thus, we achieved 99.7% coverage of the chromosome. We also sequenced a small contig of 281,116 bp on the p arm of chromosome 21.

We estimated the accuracy of the final sequence by comparing 18 overlapping sequence portions spanning 1.2 Mb. We estimate from this external checking exercise that the accuracy of the entire sequence exceeds 99.995%.

Sequence variations. Twenty-two overlapping sequence portions comprising 1.36 Mb and spread over the entire chromosome were compared for sequence variations and small deletions or insertions. We detected 1,415 nucleotide variations and 310 small deletions or insertions and confirmed them by inspecting trace files. There was an average of one sequence difference for each 787 bp, but the observed sequence variations were not evenly distributed along 21q. In the telomeric portion (21q22.3–qter) the average was one difference for each 500 bp. The highest sequence variation (one difference in 400 bp) was found in a 98-kb segment from this region. In the proximal portion (21q11–q22.3) we found on average one

difference per 1,000 bp; the lowest level was 1 in 3,600 bp in a 61-kb segment of 21q22.1.

Interspersed repeats. Table 1 summarizes the repeat content of chromosome 21. Chromosome 21 contains 9.48% Alu sequences and 12.93% LINE1 elements, in contrast with chromosome 22 which contains 16.8% Alu and 9.73% LINE1 sequences¹⁰.

Gene catalogue

The gene catalogue of chromosome 21 contains known genes, novel putative genes predicted *in silico* from genomic sequence analysis

Figure 1 The sequence map of human chromosome 21. Sequence positions are indicated in Mb. Annotated features are shown by coloured boxes and lines. The chromosome is oriented with the short p-arm to the left and the long q-arm to the right. Vertical grey box, centromere. The three small clone gaps are indicated by narrow grey vertical boxes (in proportion to estimated size) on the right of the q-arm. The cytogenetic map was drawn by simple linear stretching of the ISCN 850-band, Giemsa-stained ideogram to match the length of the sequence: the boundaries are only indicative and are not supported by experimental evidence. In the mapping phase, information on STS markers was collected from publicly available resources. The progress of mapping and sequencing was monitored using a sequence data repository in which sequences of each clone were aligned according to their map positions. A unified map of these markers was automatically generated (<http://hgp.gsc.riken.go.jp/marker/>) and enabled us to carry out simultaneous sequencing and library screening among centres. Vertical lines: markers, according to sequence position, from GDB (black; <http://www.gdb.org/>), the GB4 radiation hybrid map (blue; Whitehead Institute, Massachusetts Institute of Technology)⁴³, the G3 radiation hybrid map (dark green; Stanford Human Genome Centre, California)⁴⁴ and two linkage maps (red; Genethon; CHLC)^{45,46}. Only marker distribution is presented here: additional details, such as marker names and positions, can be found on our web sites. The *NotI* physical map of chromosome 21 was also used⁷ (*NotI* sites, light green). Genes are indicated as boxes or lines according to strand along the upper scale in three categories: known genes (category 1, red), predicted genes (categories 2 and 3, light green; category 4, light blue) and pseudogenes (category 5, violet). For genes of categories 1, 2, 3 and 5, the approved symbols from the HUGO nomenclature committee are used. CpG islands are olive (they were identified when they exceeded 400 bp in length, contained more than 55% GC, showed an observed over expected CpG frequency of >0.6 and had no match to repetitive sequences). The G+C content is shown as a graph in the middle of the Figure. It was calculated on the basis of the number of G and C nucleotides in a 100-kb sliding window in 1-kb steps across the sequence. The clone contig consists of all clones that were sequenced to 'finished' quality from all five centres in the consortium. Clones are indicated as coloured boxes by centre: red, RIKEN; dark blue, IMB; light blue, Keio; yellow, GBF; and green, MPIMG. Clones that were only partially sequenced have grey boxes on either end to show the actual or estimated clone end position. Four whole-genome libraries (RPCI-11 BAC, Keio BAC, Caltech BAC and RPCI1, 3-5 PAC) and nine chromosome-specific libraries (CMB21-BAC, Roizes-BAC, CMP21-P1, CMC21-cosmid, LLNC021, KU21D, ICRF102 and ICRF103 cosmid, and CMF21-fosmid) were used to isolate clones (see <http://hgp.gsc.riken.go.jp> or <http://chr21.r2-berlin.mpg.de> for library information). Breakpoints from chromosomal rearrangements are shown as coloured boxes according to their classification: natural (green), spontaneously occurring in cell lines (yellow), radiation induced (purple) and combinations of the above (black). Blue boxes, intra-chromosomal duplications; green boxes, inter-chromosomal duplications (see text). Alu (red) and LINE1 (blue) interspersed repeat element densities are shown in the bottom graph as the percentage of the sequence using the same method of calculation as for G+C content. The final non-redundant sequence was divided into 340-kb segments (grey boxes), with 1-kb overlaps (to avoid splitting of most exons in both segments), and has been registered, along with biological annotations, in the DDBJ/EMBL/GenBank databases under accession numbers AP001656–AP001761 (DDBJ) and AL163201–AL163306 (EMBL). Segments for the three clone gaps (accession numbers AP001742/AL163287, AP001744/AL163289 and AP001750/AL163295) have also been deposited in the databases with a number of Ns corresponding to the estimated gap lengths. The sequences and additional information can be found from the home pages of the participating centres of the chromosome 21 sequencing consortium (RIKEN, <http://hgp.gsc.riken.go.jp/>; IMB, <http://genome.imb-jena.de/>; Keio, <http://dmb.med.keio.ac.jp/>; GBF, <http://www.genome.gbf.de/>; MPI, <http://chr21.r2-berlin.mpg.de/>).

Table 1 The content of interspersed repeats in human chromosome 21

Repeat type	Total number of elements	Coverage (bp)	Coverage (%)
SINEs	15,748	3,667,752	10.84%
ALUs	12,341	3,208,437	9.48%
MIRs	3,407	459,315	1.36%
LINEs	12,723	5,245,516	15.51%
LINE1	8,982	4,372,851	12.93%
LINE2	3,741	872,665	2.58%
LTR elements	9,598	3,116,881	9.21%
MaLRs	5,379	1,646,297	4.87%
Retroviral	2,115	760,119	2.25%
MER4 group	1,396	479,451	1.42%
Other LTR	708	231,014	0.68%
DNA elements	3,950	812,031	2.40%
MER1 type	2,553	460,769	1.36%
MER2 type	851	257,653	0.76%
Mariners	168	26,235	0.08%
Other DNA elements	378	67,374	0.20%
Unclassified	64	15,234	0.05%
Total interspersed repeats	42,083	12,857,414	38.01%
Simple repeats	5,987	427,755	1.26%
Low complexity	5,868	249,449	0.74%
Total	54,045	13,551,271	40.06%
Total sequence length	33,827,477		
G+C%	40.89%		

and pseudogenes. The catalogue was arbitrarily divided into five main hierarchical categories (see below) to distinguish known genes from pure gene predictions, and also anonymous complementary DNA sequences from those exhibiting similarities to known proteins or modular domains.

The criteria governing the gene classification were based on the results of the integrated results of computational analysis using exon prediction programs and sequence similarity searches. We applied the following parameters: (1) Putative coding exons were predicted using GRAIL, GENSCAN and MZEF programs. Consistent exons were defined as those that were predicted by at least two programs. (2) Nucleotide sequence identities to expressed sequence tags (ESTs) (as identified by using BlastN with default parameters) were considered as a hallmark for gene prediction only if these ESTs were spliced into two or more exons in genomic DNA, and showed greater than 95% identity over the matched region. These criteria are conservative and were chosen to discard spurious matches arising from either cDNAs primed from intronic sites or repetitive elements frequently found in 5' or 3' untranslated regions. (3) Amino-acid similarities to known proteins or modular functional domains were considered to be significant when an overall identity of greater than 25% over more than 50 amino-acid residues was observed (as detected using BlastX with Blossum 62 matrix against the non-redundant database).

Gene categories. The results of sequence analysis were visually inspected to locate known genes, to identify new genes and to unravel novel putative transcription units after assembling consistent predicted exons into so-called *in silico* gene models. These gene predictions were also evaluated by incorporating information provided by EST and protein matches. Each gene was assigned to one of the following sub-categories:

Category 1: Known human genes (from the literature or public databases). *Subcategory 1.1:* Genes with 100% identity over a complete cDNA with defined functional association (for example, transcription factor, kinase). *Subcategory 1.2:* Genes with 100% identity over a complete cDNA corresponding to a gene of unknown function (for example, some of the KIAA series of large cDNAs).

Category 2: Novel genes with similarities over essentially their total length to a cDNA or open reading frame (ORF) of any organism. *Subcategory 2.1:* Genes showing similarity or homology to a characterized cDNA from any organism (25–100% amino-acid identity). This class defines new members of human gene families, as well as new human homologues or orthologues of genes from yeast, *Caenorhabditis elegans*, *Drosophila*, mouse and so on. *Subcategory 2.2:* Genes with similarity to a putative ORF predicted *in silico* from the genomic sequence of any organism but which currently lacks experimental verification.

Category 3: Novel genes with regional similarities to confined protein regions. *Subcategory 3.1:* Genes with amino-acid similarity confined to a protein region specifying a functional domain (for example, zinc fingers, immunoglobulin domains). *Subcategory 3.2:* Genes with amino-acid similarity confined to regions of a known protein without known functional association.

Category 4: Novel anonymous genes defined solely by gene prediction. These are putative genes lacking any detectable similarity to known proteins or protein motifs. These models are based solely on spliced EST matches, consistent exon prediction or both.

Subcategory 4.1: Predicted genes composed of a pattern of two or more consistent exons (located within <20 kb) and supported by spliced EST match(es). *Subcategory 4.2:* Predicted genes corresponding to spliced EST(s) but which failed to be recognized by exon prediction programs. *Subcategory 4.3:* Predicted genes composed only of a pattern of consistent exons without any matches to EST(s) or cDNA. Intuitively, predicted genes from subcategory 4.1 are considered to have stronger coding potential than those of subcategory 4.3.

Category 5: Pseudogenes may be regarded as gene-derived DNA sequences that are no longer capable of being expressed as protein products. They were defined as predicted polypeptides with strong similarity to a known gene, but showing at least one of the following features: lack of introns when the source gene is known to have an intron/exon structure, occurrence of in-frame stop codons, insertions and/or deletions that disrupt the ORF or truncated matches. Generally, this was an unambiguous classification.

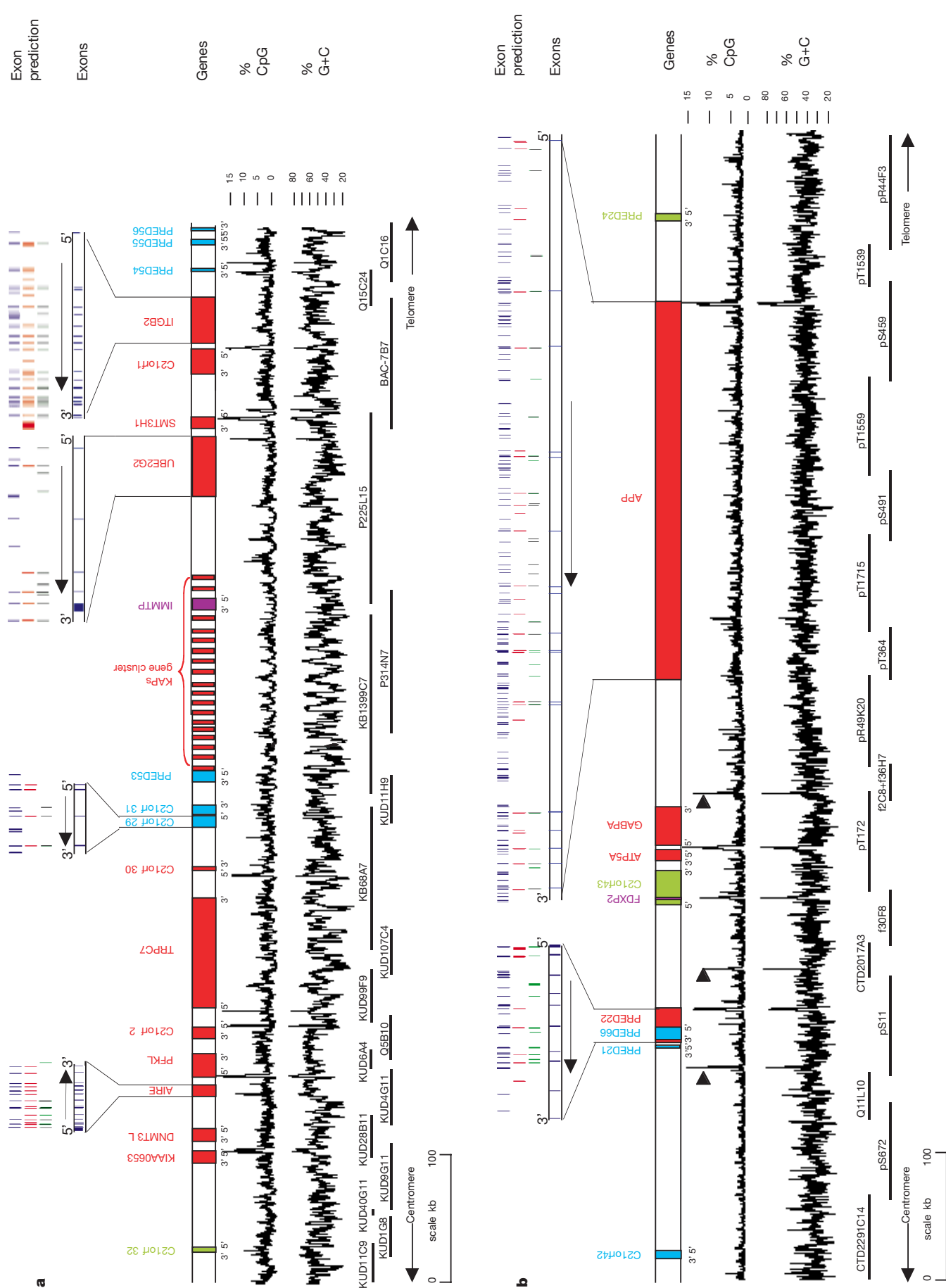
When a gene could fulfil more than one of these criteria, it was placed into the higher possible category (for example, gene prediction with spliced EST exhibiting a significant match to a known protein was placed in subcategory 2.2 rather than 4.2).

The gene content of chromosome 21. For the gene catalogue of chromosome 21, see Table 2. The chromosome contains 225 genes and 59 pseudogenes. Of these, 127 correspond to known genes (subcategories 1.1 and 1.2) and 98 represent putative novel genes predicted *in silico* (categories 2, 3 and 4). Of the novel genes, 13 are similar to known proteins (subcategories 2.1 and 2.2), 17 are anonymous ORFs featuring modular domains (subcategories 3.1 and 3.2), and most (68 genes) are anonymous transcription units with no similarity to known proteins (subcategories 4.1, 4.2 and 4.3). Our data show that about 41% of the genes that were identified on chromosome 21 have no functional attributes.

In a rough generic description, the gene catalogue of chromosome 21 contains at least 10 kinases (PRED1, PRSS7, C21orf7, PRED33, PRKCBP2, DYRK1A, ANKDR3, SNF1LK, PDXK and PFKL), five genes involved in ubiquitination pathways (USP25, USP16, UBASH, UBE2G2 and SMT3H1), five cell adhesion molecules (NCAM2, IGSF5, C21orf43, DSCAM and ITGB2), a number of transcription factors and seven ion channels (C21orf34, KCNE2, KCNE1, CILC1L, KCNJ6, KCNJ15 and TRPC7). Several clusters of functionally related genes are arranged in tandem arrays on 21q, indicating the likelihood of ancient sequential rounds of gene duplication. These clusters include the five members of the interferon receptor family that spans 250 kb on 21q (positions 20,179,027–20,428,899), the trefoil peptide cluster (TFF1, TFF2 and TFF3) spanning 54 kb on 21q22.3 (positions 29,279,519–29,333,970) and the keratin-associated protein (KAP) cluster spanning 164 kb on 21q22.3 (positions 31,468,577–31,632,094) (Table 2). The last contains 18 units of this highly repetitive gene family featuring genes and different pseudogene fragments and revealing inverted duplications within the gene cluster (described below). Finally, the p arm of chromosome 21 contains at least one gene (TPTE) encoding a putative tyrosine phosphatase. This is the first description of a protein-coding gene mapping to the p arm of an acrocentric chromosome. However, the functional activity of this gene remains to be demonstrated.

Chromosome 21 contains a very low number of identified genes (225) compared with the 545 genes reported for chromosome 22 (ref. 10). Figure 1 shows the overall distribution of the 225 genes and 59 pseudogenes on chromosome 21 in relation to compositional features such as G+C content, CpG islands, Alu and L1 repeats and the positions of selected STSs, polymorphic markers and chromosomal breakpoints. Earlier reports indicated that gene-rich regions are Alu rich and LINE1 poor, whereas gene-poor regions contain more LINE1 elements at the expense of Alu sequences¹¹. Our data, and the comparison with chromosome 22, support these findings (see Tables 1 and 2, Fig. 1 and ref. 10). There is a large 7-Mb region

Table 2 Gene catalogue of chromosome 21. The table displays the gene symbol, accession number, gene description, gene category, orientation, gene start position, gene end position, genomic size and corresponding genomic clone name. The gene categories are colour coded as follows: known genes (category 1) in red, novel genes with similarities to characterized cDNAs from any organism and novel genes with similarities to protein domains (categories 2 and 3) in green, novel gene prediction (category 4) in blue, and pseudogenes (category 5) in purple. Coordinates are given in base pairs.



(between 5 and 12 Mb on Fig. 1) with low G+C content (35% compared with 43% for the rest of the chromosome) that correlates with a paucity of both Alu sequences and genes. Only two known genes (PRSS7 and NCAM2) and five predicted genes can be found in this region. Further reinforcing the concept that compositional features correlate with gene density, Fig. 2 compares the genomic organization and gene density in a 831-kb G+C-rich DNA region (53%; Fig. 2a) with that of a 915-kb DNA stretch representative of a G+C-poor region (39.5%; Fig. 2b). Figure 2a shows eleven known genes, seven predicted genes, one pseudogene and the KAP cluster. Figure 2b shows four known genes, five predicted genes and one pseudogene. Figure 2 also displays examples of exon/intron structures as defined by the exon prediction programs in parallel with the real gene structure that was obtained by sequence alignment using the cognate mRNA. Most exons were predicted by the combination of the three programs. However, MZEF tends to overpredict exons compared with GRAIL and GENSCAN, in particular for the large APP gene. In addition, CpG islands correlate well as indicators of the 5' end of genes in both of these regions.

Structural features of known and predicted genes. Among the 127 known genes, 22 genes are larger than 100 kb, the largest being DSCAM (840 kb). Seven of the largest known genes cover 1.95 Mb and lie within a region of 4.5 Mb (positions 23.7 Mb–28.2 Mb) that contains only four predicted genes and two pseudogenes. The average size of the genes is 39 kb, but there is a bias in favour of the category 1 genes. Known genes have a mean size of 57 kb, whereas predicted genes (categories 2, 3 and 4) have a mean size of 27 kb. This is not unexpected, because of the inherent difficulties in extending exon prediction to full-length gene identification. For instance, exon prediction and EST findings are usually not exhaustive. This would also explain the fact that 69% of the predicted genes have no similarity to known proteins.

Despite the shortcomings of current gene prediction methods, all known genes previously shown to map on chromosome 21 (ref. 12) were identified independently by *in silico* methods. Patterns of consistent exon prediction alone were sufficient to locate at least partial gene structures for more than 95% of these. This was true even for large A+T-rich genes, such as NCAM2, APP (Fig. 2b) and GRIK1. These three genes are several hundred kilobases long with a G+C content of 38–40%, but most exons were well predicted and enough introns were sufficiently small that a clear pattern of consistent exons was seen. In addition, more than 95% of the known genes were independently identified from spliced ESTs. Characteristics of genes that could be missed using our detection methods include those with poor exon prediction and long 3' untranslated regions (>2 kb); those with poor exon prediction and very restricted expression pattern; and those with very large introns (>30 kb).

We designed our gene identification criteria to extract most of the coding potential of the chromosome and to minimize false positive predictions. Errors to be expected in the predictions include false positive exons, incorrect splice sites, false negative exons, fusion of multiple genes into one transcription unit and separation of a single gene into two or more transcription units. We believe that our method is sufficiently robust to pinpoint real genes, but our models still require experimental validation. In a pilot experiment on 14 predicted category 4 genes we performed RT-PCR (PCR with reverse transcription) in 12 tissues. We could confirm 11 genes and connect two gene predictions into a single transcription unit.

Pseudogenes are often overlooked in a gene catalogue aimed at specifying functional proteins, but they may be important in influencing recombination events. The 59 pseudogenes described here are not randomly located in the chromosome (Fig. 1). Twenty-four pseudogenes are distributed in the first 12 Mb of 21q, which is a gene-poor region. In contrast, a cluster of 11 pseudogenes was found within a 1-Mb stretch of DNA that is gene rich and corresponds precisely to the highest density of Alu sequences on the chromosome (positions 22,421,026–23,434,597).

Base composition and gene density. It is tempting to speculate on possible correlations between the base composition, gene density and molecular architecture of the chromosome bands. Giemsa-dark chromosomal bands are comprised of L isochores (<43% G+C), whereas Giemsa-light bands have variable composition. The latter include L, H1/H2 (43–48% G+C) and H3 isochores (>48% G+C)¹³. In humans, the average gene density is around one gene per 150 kb in L, one per 54 kb in H1/H2 and one per 9 kb in H3 isochores¹⁴. The proximal half of 21q (from 0.2 to 17.7 Mb of Fig. 1), which corresponds mainly to the large Giemsa dark band, 21q21, comprises a long continuous L isochore, harbouring extensive stretches of 34–37% G+C, and rare segments of more than 40% G+C. Twenty-five category 1 genes and 33 category 2–4 genes were found in this region, giving an average density of one gene per 301 kb.

The distal half of 21q (17.7–33.5 Mb) largely comprises stretches of H1/H2 isochores alternating with L isochores, and H3 isochores localized within the region spanning positions 29–33.5 Mb. The overall gene density in the telomeric half is much higher than that in the proximal half: 101 genes of category 1 and 66 genes of categories 2–4 were found in this region, giving an average of about one gene per 95 kb. The DSCAM gene, found within an L isochore in this region, spans 834 kb. In contrast, the region spanning the H3 isochores contains 46 category 1 genes and 31 category 2–4 genes, averaging one gene per 58 kb.

The L isochores have lower gene density than that predicted from whole-genome analysis: one gene per 301 kb compared with one per 150 kb. The H3 isochores are also lower in gene content, averaging one gene per 58 kb compared with one gene per 9 kb estimated for the genome as a whole. This discrepancy may be due to an overestimation of the total number of human genes based on EST data (see below). Alternatively, we may have missed half of the genes on this chromosome. This second possibility is unlikely as more than 95% of the known genes have been predicted using our criteria.

Chromosomal structural features

Duplications within chromosome 21. The unmasked sequence of the whole chromosome was compared with itself to detect intra-chromosomal duplications. We identified a 10-kb duplication in the pericentromeric regions of the p- and q-arms (Fig. 3a). The p-arm copy extends from 190 to 199 kb of the p-arm contig, and the q-arm copy extends from 405 to 413 kb of the 21q sequence. We identified a CpG island on the centromeric side of the duplication in the p-arm, indicating that there may be an active gene in the vicinity of the duplicated regions. A similar structure was reported for chromosome 10 (ref. 15), so such repeats close to the centromere may have a functional role. The pericentromeric region in the q-arm also contains several duplications, including several clusters of α -satellite sequences and even telomeric satellites.

Another duplication corresponding to a large 200-kb region has been identified in proximal and distal locations on 21q (Fig. 3b). This duplication was previously reported¹⁶ but was not analysed in detail at the sequence level. The proximal copy is located from 188 to 377 kb in 21q11.2, whereas the distal copy lies in 21q22 and extends from 14,795 to 15,002 kb. The two copies are highly conserved and show 96% identity. We detected two large inversions, several other rearrangements and several translocations or duplications within the duplicated units (Fig. 3b), which caused segmentation of the

Figure 2 Gene organization on chromosome 21. **a**, A G+C-rich region of the telomeric part; **b**, an AT-rich region of the centromeric part. Genes are represented by coloured boxes. Category 1, red; categories 2 and 3, green; category 4, blue; category 5, violet. Predicted exons shown in the enlarged gene areas are represented as: MZEF, blue; GenScan, red; Grail, green. Arrowheads, orphan CpG islands that may indicate the presence of a cryptic gene.

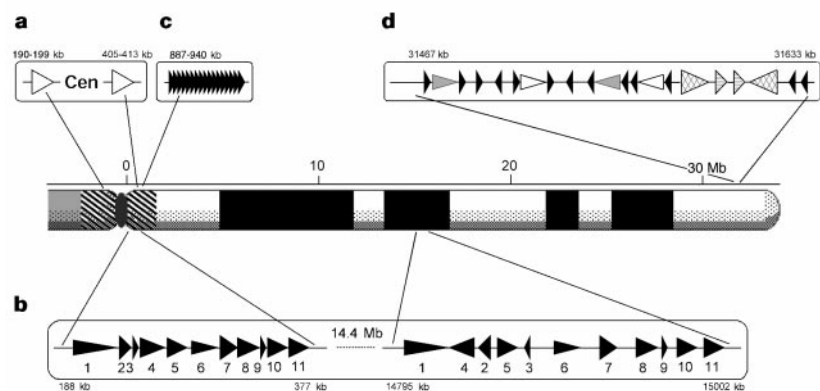


Figure 3 Schematic view of the duplicated regions in chromosome 21 as described in the text. **a–d**, Duplicated regions. The positions of each repeat structure are shown in kb starting at the centromere. The arrowheads represent the orientation and approximate size of each repetitive unit.

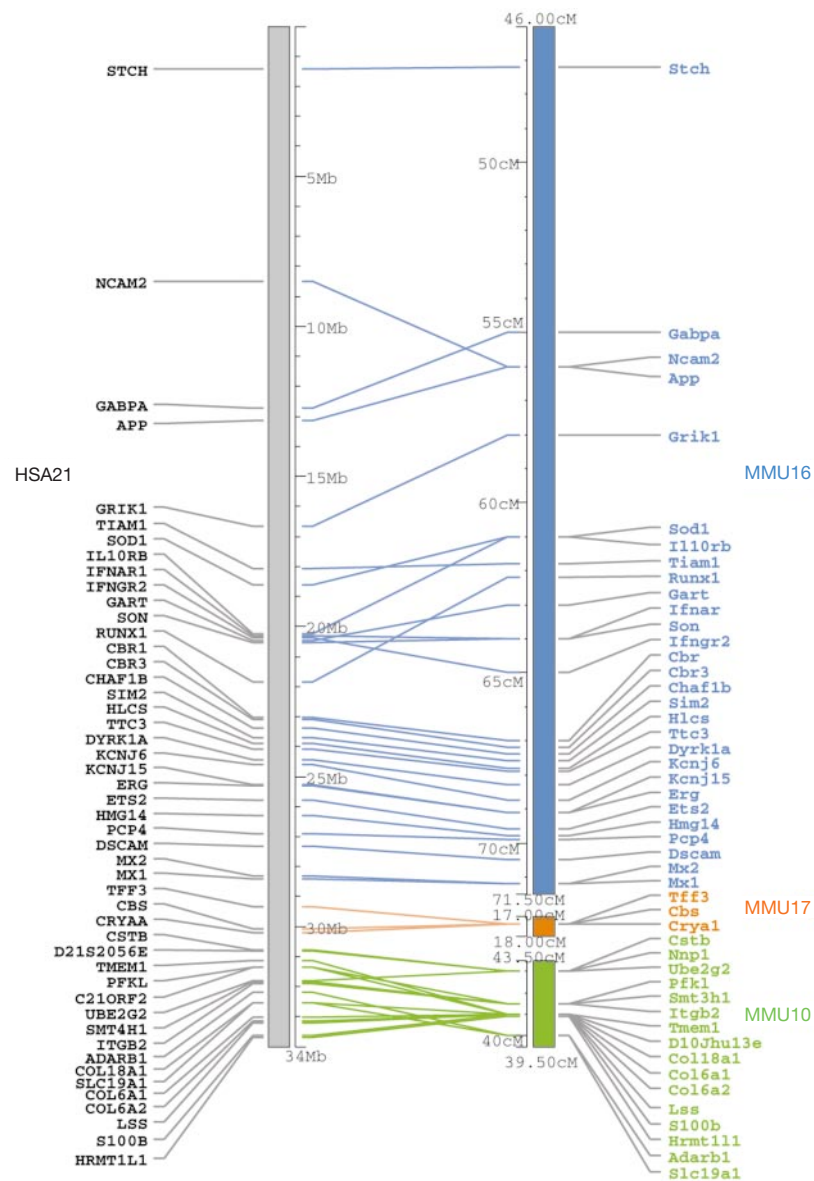


Figure 4 Schematic view of the syntenic regions between human chromosome 21 (HSA21) and mouse chromosomes 16 (MMU16), 17 (MMU17) and 10 (MMU10). Left: sequence map of human chromosome 21. Right: corresponding mouse chromosomes. Each pair of syntenic markers is joined with a line.

units into at least 11 pieces. The distal copy is 207 kb long and the proximal copy is 189 kb; the 18-kb size difference between the two duplicated segments is due to insertions in the distal copy, deletions in the proximal copy or both.

In the region on 21q between 887 and 940 kb a block of sequence is repeated 17 times (Fig. 3c). The similarity of these repetitive units indicates that they were formed by a recent triplication event of a region of six repeat unit blocks, which had in turn been generated by duplication of a three-block unit.

Another repeat sequence lies between the TRPC7 and UBE2G2 genes on 21q22.3 (31,467–31,633 kb). This feature corresponds to the 166-kb KAP gene and pseudogene cluster described above (Fig. 2a). A 0.5–1-kb segment is repeated at least 13 times, with 5–10-kb spacer intervals (Fig. 3d). The repeat units share more than 91% identity with each other.

Comparison of chromosome 21 with chromosome 22. The two chromosomes are similar in size, and both are acrocentric. The gene density, however, is much higher on chromosome 22 (ref. 10). We detected sequence similarity in the pericentromeric and sub-telomeric regions of both chromosomes. For example, two different regions in the 21p contig (42–84 kb; 239–263 kb) are duplicated in 22q (1043–1067 kb; 1539–1564 kb). These duplications are located within the pericentromeric regions of both chromosomes¹⁷. Half of the first region is further duplicated at the position 22,223–22,248 kb in chromosome 22. In addition, two inverted duplications in 21q at 88–156 kb and 646–751 kb have also been observed on 22q at positions 572–637 kb and 45–230 kb. Large clusters of α -satellite sequences (10 kb for chromosome 21 and 119 kb for chromosome 22) are located on 21q (88–156 kb) and 22q (572–637 kb).

The most telomeric clone, F50F5, isolated from the chromosome-specific CMF21 fosmid library, contains a telomeric repeat array that represents the hallmark of the telomeric end of a chromosome. This array was missing in the chromosome 22q sequence¹⁰. However, the 22q sequence ends very near to the telomere, considering that it shows strong homology with a 2.5–10-kb stretch of telomeric sequence present in F50F5.

Comparison of chromosome 21 with other autosomes. In the most telomeric region of chromosome 21 we also identified a novel repeat structure featuring a non-identical 93-bp unit that is repeated 10 times. This block of 93-bp repeats is located 7.5 kb from the start point of the telomeric array. Similar 93-bp repeat sequences were also detected by BLAST analysis in chromosomes 22, 10 and 19. FISH analysis data suggest that this 93-bp repeat unit is also located on 5qter, 7pter, 17qter, 19pter, 19qter, 20pter, 21qter and 22qter, as well as on other chromosomal ends. Thus, this 93-bp repeat may be a common structural feature shared by many human telomeres.

We have found some paralogous regions between chromosome 21 and other human chromosomes, which were also pointed out by metaphase FISH analysis of the corresponding genomic clones. For example, a 100-kb region of clone B15L0C0 located on 21p is shared with chromosomes 4, 7, 20 and 22. A second homologous region of 50 kb on 21q between 15,530 and 15,580 kb is shared with a segment on chromosome 16 between the genes 44M2.1 and 44M2.2. More details on these regions can be found at <http://hgp.gsc.riken.go.jp/>.

Syntenies with mouse. Human chromosome 21 shows conserved syntenies to mouse chromosomes 16, 17 and 10 (<http://www.informatics.jax.org/>). Figure 4 shows a comparative map of human chromosome-21-specific genes with their mouse orthologues. A number of inversions can be seen. These changes in gene order may be due to rearrangements during genome evolution. Alternatively, they may reflect the fact that the mouse gene map is still inaccurate because it is based on linkage and physical mapping.

Breakpoints. Figure 1 shows the locations of 39 breakpoints on the physical map. Here we describe several classes of breakpoint, all of which either occurred naturally in the human population before hybrid construction or were induced by irradiation. The natural

breakpoints arose mainly from reciprocal translocations of chromosome 21 with other human chromosomes (6;21, 4;21, 3;21, 1;21, 8;21, 10;21, 11;21 and 21;22). A second class of naturally occurring breakpoints derived from intrachromosomal rearrangements of chromosome 21 (ACEM, 6918, MRC2, R210 and DEL21). A third class of breakpoints, designated 3x1, 3x2, 1x4D, 1x4F and 1x18, were generated experimentally by irradiation of hybrids containing intact chromosome 21q arms¹⁸. Hybrids 2Fur, 750 and 511 represent rearrangements of chromosome 21 that occurred spontaneously in somatic cell hybrids. All of these chromosome derivatives were isolated in Chinese hamster ovary (CHO) $\times$ human somatic cell hybrids.

Fine mapping revealed an uneven distribution of breakpoints that fell roughly in two clusters on chromosome 21. Nine breakpoints occur within the pericentromeric region (0–2.2 Mb) and another nine are located within a 2.4-Mb region in 21q22 (20.1–22.5 Mb) (Fig. 1). In contrast, large regions are totally devoid of breakpoints. For instance, only two translocation breakpoints are located in the 10-Mb region between 4.95 and 14.4 Mb of the q arm.

Several breakpoints occur within or near the duplicated regions described above. For instance, three breakpoints (1x4D, 1x18 and 2Fur) occur between positions 100 and 400 kb on 21q. This region corresponds to the proximal copy of the large duplicated region described in Fig. 3b. Another breakpoint (ACEM) occurs between positions 14,400 and 14,525 kb, close to the distal copy of this duplicated region. We also found a naturally occurring 21;22 translocation breakpoint (position 31,350–31,380 kb) in the KAP cluster.

Duplicated regions may mediate certain mechanisms involved in chromosomal rearrangement. It is likely that similar sequence features may be important for duplication, genetic recombination and chromosomal rearrangement. Further sequence analysis will help to unravel the underlying molecular mechanisms of chromosome breakage and recombination.

Recombination. The distribution of the recombination frequency on chromosome 21 is different in males and females¹². In Fig. 5 genetic distances of known polymorphic markers from male, female and sex-average maps are compared with the distances in nucleotides on 21q. The recombination frequency is relatively higher near the centromere in females and near the telomere in males. This confirms earlier analysis based on physical maps¹¹. Unlike chromosome 22, chromosome 21 does not appear to contain particular regions with a steep increase in recombination frequency in the middle of the chromosome.

Medical implications

Down syndrome. Besides the constant feature of mental retardation, individuals with Down syndrome also frequently exhibit congenital heart disease, developmental abnormalities, dysmorphic features, early-onset Alzheimer's disease, increased risk for specific leukaemias, immunological deficiencies and other health problems¹⁹. Ultimately, all these phenotypes are the result of the presence of three copies of genes on chromosome 21 instead of two. Data from transgenic mice indicate that only a subset of the genes on chromosome 21 may be involved in the phenotypes of Down syndrome²⁰. Although it is difficult to select candidate genes for these phenotypes, some gene products may be more sensitive to gene dosage imbalance than others. These may include morphogens, cell adhesion molecules, components of multi-subunit proteins, ligands and their receptors, transcription regulators and transporters. The gene catalogue now allows the hypothesis-driven selection of different sets of candidates, which can then be used to study the molecular pathophysiology of the gene dosage effects. The complete catalogue will also provide the opportunity to search systematically for candidate genes without pre-existing hypotheses.

Monogenic disorders. Mutations in 14 known genes on chromo-

some 21 have been identified as the causes of monogenic disorders including one form of Alzheimer's disease (APP), amyotrophic lateral sclerosis (SOD1), autoimmune polyglandular disease (AIRE), homocystinuria (CBS) and progressive myoclonus epilepsy (CSTB); in addition, a locus for predisposition to leukaemia (AML1) has been mapped to 21q (for details of each of these disorders, see <http://www.ncbi.nlm.nih.gov/omim/>). The cloning of some of these genes, including the AIRE gene^{21,22}, was facilitated by the sequencing effort. Loci for the following monogenic disorders have not yet been cloned: recessive nonsyndromic deafness (DFNB10 (ref. 23) and DFNB8 (ref. 24)), Usher syndrome type 1E²⁵, Knobloch syndrome²⁶ and holoprocencephaly type 1 (HPE1 (ref. 27)). The gene catalogue and mapping coordinates will help in their identification. Mutation analysis of candidate genes in patients will lead to the cloning of the responsible genes.

Complex phenotypes. Two loci conferring susceptibility to complex diseases have been mapped to chromosome 21 (one for bipolar affective disorder²⁸ and one for familial combined hyperlipidaemia²⁹) but the genes involved remain elusive.

Neoplasias. Loss of heterozygosity has been observed for specific regions of chromosome 21 in several solid tumours^{30–36} including cancers of the head and neck, breast, pancreas, mouth, stomach, oesophagus and lung. The observed loss of heterozygosity indicates that there may be at least one tumour suppressor gene on this chromosome. The decreased incidence of solid tumours in individuals with Down syndrome indicates that increased dosage of some chromosome 21 genes may protect such individuals from these tumours^{37–39}. On the other hand, Down syndrome patients have a markedly increased risk of childhood leukaemia¹⁹, and trisomy of chromosome 21 in blast cells is one of the most common chromosomal aneuploidies seen in childhood leukaemias⁴⁰.

Chromosome abnormalities. Chromosome 21 is also involved in chromosomal aberrations including monosomies, translocations and other rearrangements. The availability of the mapped and sequenced clones now provides the necessary reagents for the accurate diagnosis and molecular characterization of constitutional and somatic chromosomal abnormalities associated with various phenotypes. This, in turn, will aid in identifying genes involved in mechanisms of disease development.

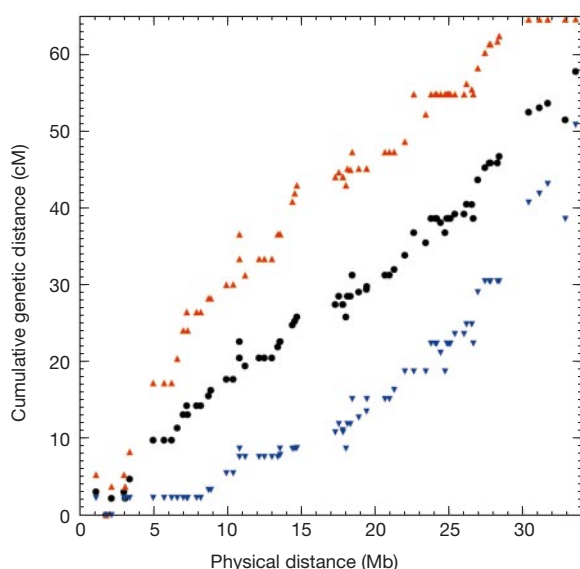


Figure 5 Comparison of the genetic map and the sequence map of chromosome 21 aligned from centromere to telomere. Genetic distance in cM; physical distance in Mb. Each spot reflects the position of a particular genetic marker retrieved from <http://www.marshmed.org>. Black circles, sex-average; orange upwards triangles, female; blue downwards triangles, male.

The analysis of the genetic variation of many of the genes on chromosome 21 is of particular importance in the search for associations of polymorphisms with complex diseases and traits. Single nucleotide polymorphism (SNP) genotyping may also aid in the identification of modifier genes for numerous pathologies. Similarly, SNPs are useful tools in the development of diagnostic and predictive tests, which may eventually lead to individualized treatments. Chromosome-21-specific nucleotide polymorphisms will also facilitate evolutionary studies.

Discussion

Our sequencing effort provided evidence for 225 genes embedded within the 33.8 Mb of genomic DNA of chromosome 21. Five hundred and forty-five genes have been identified in the 33.4 Mb of chromosome 22 (ref. 10). These data support the conclusion that chromosome 22 is gene-rich, whereas chromosome 21 is gene-poor. This finding is in agreement with data from the mapping of 30,181 randomly selected Unigene ESTs⁴¹. These two chromosomes together represent about 2% of the human genome and collectively contain 770 genes. Assuming that both chromosomes combined reflect an average gene content of the genome, we estimate that the total number of human genes may be close to 40,000. This figure is considerably lower than previous estimates, which range from 70,000 to 140,000 (ref. 42), and which were mainly based on EST clustering. It is possible that not all of the genes on chromosomes 21 and 22 have been identified. Alternatively, our assumption that the two chromosomes represent good models may be incorrect.

Our analysis of the chromosomal architecture revealed repeat units, duplications and breakpoints. A 93-bp repeat in the telomeric region, which was also found in other chromosomes, should provide a basis for studying the structural and functional organization and evolution of the telomere. One striking feature of chromosome 21 is that there is a 7-Mb region (positions 5.5–12.5 Mb) that contains only one gene. This region is much larger than the whole genome of *Escherichia coli*, but the evolutionary process permitted the existence of such a gene-poor DNA segment. Three other 1-Mb regions on 21q are also devoid of genes. Together, these gene-poor regions comprise almost 10 Mb, which is one-third of chromosome 21. Chromosome 22 also has a 2.5-Mb region near the telomeric end, as well as two other regions, each of 1 Mb, which are devoid of genes. We propose that similar large gene-less or gene-poor regions exist in other mammalian chromosomes. These regions may have a functional or architectural significance that has yet to be discovered.

Having the complete contiguous sequence of human chromosomes will change the methodology for finding disease-related genes. Disease genes will be identified by combining genetic mapping with mutation analysis in positional candidate genes. The laborious intermediate steps of physical mapping and sequencing are no longer necessary. Therefore, any individual investigator will be able to participate in disease gene identification.

The complete sequence analysis of human chromosome 21 will have profound implications for understanding the pathogenesis of diseases and the development of new therapeutic approaches. The clone collection represents a useful resource for the development of new diagnostic tests. The challenge now is to unravel the function of all the genes on chromosome 21. RNA expression profiling with all chromosome-21-specific genes may allow the identification of up- and downregulated genes in normal and disease samples. This approach will be particularly important for studying expression differences in trisomy and monosomy 21. Furthermore, chromosome-21-homologous genes can be systematically studied by overexpression and deletion in model organisms and mammalian cells.

The relatively low gene density on chromosome 21 is consistent with the observation that trisomy 21 is one of the only viable human autosomal trisomies. The chromosome 21 gene catalogue will open

new avenues for deciphering the molecular bases of Down syndrome and of aneuploidies in general. □

Methods

Details of the protocols used by the five sequencing centres are available from our web sites (see below), including methods for the construction of sequence-ready maps and for sequencing large insert clones by shotgun cloning and nested deletion. Many software programs were used by the five groups for data processing, sequence analysis, gene prediction, homology searches, protein annotation and searches for motifs using pfam and SMART. Most of these programs are in the public domain. Software suites have been developed by the consortium members to allow efficient analysis. All information is available from the following web pages: RIKEN: <http://hgp.gsc.riken.go.jp>; Institut für Molekulare Biotechnologie, Jena: <http://genome.imb-jena.de>; Keio University: <http://www.dmb.med.keio.ac.jp>; GBF-Braunschweig: <http://genome.gbf.de>; Max-Planck-Institut für Molekulare Genetik (MPIMG), Berlin: <http://chr21.rz-berlin.mpg.de>.

Received 17 April; accepted 3 May 2000.

1. Lejeune, J., Gautier, M. & Turpin, R. Etude des chromosomes somatique des neuf enfants mongoliens. *CR Acad. Sci. Paris* **248**, 1721–1722 (1959).
2. McInnis, M. G. *et al.* A linkage map of human chromosome 21: 43 PCR markers at average intervals of 2.5 cM. *Genomics* **16**, 562–571 (1993).
3. Chumakov, I. *et al.* Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* **359**, 380–387 (1992).
4. Nizetic, D. *et al.* An integrated YAC-overlap and "cosmid-pocket" map of the human chromosome 21. *Hum. Mol. Genet.* **3**, 759–770 (1994).
5. Gardiner, K. *et al.* YAC analysis and minimal tiling path construction for chromosome 21q. *Somat. Cell Mol. Genet.* **21**, 399–414 (1995).
6. Korenberg, J. R. *et al.* A high-fidelity physical map of human chromosome 21q in yeast artificial chromosomes. *Genome Res.* **5**, 427–443 (1995).
7. Ichikawa, H. *et al.* A *NotI* restriction map of the entire long arm of human chromosome 21. *Nature Genet.* **4**, 361–366 (1993).
8. Hildmann, T. *et al.* A contiguous 3-Mb sequence-ready map in the S3-MX region on 21q22.2 based on high-throughput nonisotopic library screenings. *Genome Res.* **9**, 360–372 (1999).
9. Hattori, M. *et al.* A novel method for making nested deletions and its application for sequencing of a 300 kb region of human APP locus. *Nucleic Acids Res.* **25**, 1802–1808 (1997).
10. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
11. Korenberg, J. R. & Rykowski, M. C. Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* **53**, 391–400 (1988).
12. Antonarakis, S. E. 10 years of Genomics, chromosome 21, and Down syndrome. *Genomics* **51**, 1–16 (1998).
13. Saccone, S. *et al.* Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl Acad. Sci. USA* **90**, 11929–11933 (1993).
14. Zoubak, S., Clay, O. & Bernardi, G. The gene distribution of the human genome. *Gene* **174**, 95–102 (1996).
15. Jackson, M. S. *et al.* Sequences flanking the centromere of human chromosome 10 are a complex patchwork of arm-specific sequences, stable duplications and unstable sequences with homologies to telomeric and other centromeric locations. *Hum. Mol. Genet.* **8**, 205–215 (1999).
16. Dutriaux, A. *et al.* Cloning and characterization of a 135- to 500-kb region of homology on the long arm of human chromosome 21. *Genomics* **22**, 472–477 (1994).
17. Ruault, M. Juxta-centromeric region of human chromosome 21 is enriched for pseudogenes and gene fragments. *Gene* **239**, 55–64 (1999).
18. Graw, S. L. *et al.* Molecular analysis and breakpoint definition of a set of human chromosome 21 somatic cell hybrids. *Somat. Cell Mol. Genet.* **21**, 415–428 (1995).
19. Epstein, C. J. in *The Metabolic and Molecular Bases of Inherited Disease* (eds Scriver, C. R. *et al.*) 749–794 (McGraw-Hill, New York, 1995).
20. Kola, I. & Hertzog, P. J. Animal models in the study of the biological function of genes on human chromosome 21 and their role in the pathophysiology of Down syndrome. *Hum. Mol. Genet.* **6**, 1713–1727 (1997).
21. Nagamine, K. *et al.* Positional cloning of the APECED gene. *Nature Genet.* **17**, 393–398 (1997).
22. The Finnish-German APECED Consortium. An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. Autoimmune Polyendocrinopathy-Candidiasis-Ectodermal Dystrophy. *Nature Genet.* **17**, 399–403 (1997).
23. Bonné-Tamir, B. *et al.* Linkage of congenital recessive deafness (Gene DFN10) to chromosome 21q22.3. *Am. J. Hum. Genet.* **58**, 1254–1259 (1996).
24. Veske, A. *et al.* Autosomal recessive non-syndromic deafness locus (DFNB8) maps on chromosome 21q22 in a large consanguineous kindred from Pakistan. *Hum. Mol. Genet.* **5**, 165–168 (1996).
25. Chaib, H. *et al.* A newly identified locus for Usher syndrome type I, USH1E, maps to chromosome 21q21. *Hum. Mol. Genet.* **6**, 27–31 (1997).
26. Sertie, A. L. *et al.* A gene which causes severe ocular alterations and occipital encephalocele (Knobloch syndrome) is mapped to 21q22.3. *Hum. Mol. Genet.* **5**, 843–847 (1996).
27. Estabrooks, L. L., Rao, K. W., Donahue, R. P., & Aylsworth, A. S. Holoprosencephaly in an infant with a minute deletion of chromosome 21(q22.3). *Am. J. Med. Genet.* **36**, 306–309 (1990).
28. Straub, R. E. *et al.* A possible vulnerability locus for bipolar affective disorder on chromosome 21q22.3. *Nature Genet.* **8**, 291–296 (1994).
29. Pajukanta, P. *et al.* Genomewide scan for familial combined hyperlipidemia genes in Finnish families, suggesting multiple susceptibility loci influencing triglyceride, cholesterol, and apolipoprotein B levels. *Am. J. Hum. Genet.* **64**, 1453–1463 (1999).
30. Sakata, K. *et al.* Commonly deleted regions on the long arm of chromosome 21 in differentiated adenocarcinoma of the stomach. *Genes Chromosome Cancer* **18**, 318–321 (1997).

31. Kohno, T. *et al.* Homozygous deletion and frequent allelic loss of the 21q11.1–q21.1 region including the ANA gene in human lung carcinoma. *Genes Chromosome Cancer* **21**, 236–243 (1998).
32. Ohgaki, K. *et al.* Mapping of a new target region of allelic loss to a 6-cM interval at 21q21 in primary breast cancers. *Genes Chromosome Cancer* **23**, 244–247 (1998).
33. Yamamoto, N. *et al.* Frequent allelic loss/imbalance on the long arm of chromosome 21 in oral cancer: evidence for three discrete tumor suppressor gene loci. *Oncol. Rep.* **6**, 1223–1227 (1999).
34. Ghadimi, B. M. *et al.* Specific chromosomal aberrations and amplification of the AIB1 nuclear receptor coactivator gene in pancreatic carcinomas. *Am. J. Pathol.* **154**, 525–536 (1999).
35. Bockmuhl, U. *et al.* Genomic alterations associated with malignancy in head and neck cancer. *Head Neck* **20**, 145–151 (1998).
36. Schwendel, A. *et al.* Chromosome alterations in breast carcinomas: frequent involvement of DNA losses including chromosomes 4q and 21q. *Br. J. Cancer* **78**, 806–811 (1998).
37. Satge, D. *et al.* M. A tumor profile in Down syndrome. *Am. J. Med. Genet.* **78**, 207–216 (1998).
38. Hasle, H., Clemmensen, I. H., & Mikkelsen, M. Risks of leukaemia and solid tumours in individuals with Down's syndrome. *Lancet* **355**, 165–169 (2000).
39. Satge, D. *et al.* A lack of neuroblastoma in Down syndrome: a study from 11 European countries. *Cancer Res.* **58**, 448–452 (1998).
40. Wan, T. S., Au, W. Y., Chan, J. C., Chan, L. C. & Ma, S. K. Trisomy 21 as the sole acquired karyotypic abnormality in acute myeloid leukemia and myelodysplastic syndrome. *Leuk. Res.* **23**, 1079–1083 (1999).
41. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
42. Fields, C., Adams, M. D., White, O. & Venter, J. C. How many genes in the human genome? *Nature Genet.* **7**, 345–346 (1994).
43. Gyapay, G. *et al.* A radiation hybrid map of the human genome. *Hum. Mol. Genet.* **5**, 339–346 (1996).
44. Stewart, E. A. *et al.* An STS-based radiation hybrid map of the human genome. *Genome Res.* **7**, 422–433 (1997).
45. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154 (1996).
46. Murray, J. C. *et al.* A comprehensive human linkage map with centimorgan density. *Science* **265**, 2049–2054 (1994).

Acknowledgements

The RIKEN group thank T. Itoh and C. Kawagoe for support of computational data management, M. Ohira and R. Ohki for clones and the members listed on <http://hgp.gsc.riken.go.jp> for technical support. The Jena group thank C. Baumgart, M. Dette, B. Drescher, G. Glöckner, S. Kluge, G. Nyakatura, M. Platzer, H.-P. Pohle, R. Schattevoi, M. Schilling, J. Weber and all present and past members of the sequencing teams. The Keio group thank E. Nakato, M. Asahina, A. Shimizu, I. Abe, J. Wang, N. Sawada, M. Tatsuyama, M. Takahashi, M. Sasaki, H. Harigai and all members of the sequencing team, past and present. The MPIMG group thank M. Klein, C. Steffens, S. Arndt, K. Heitmann, I. Langer, D. Buczek, J. O'Brien, M. Christensen, T. Hildmann, I. Szulzewski, E. Hunt and G. Teltow for technical support, and T. Haaf and A. Palotie for help with FISH. The German groups (IMB, GBF and MPIMG) thank the Resource Center of the German Human Genome Project (RZPD) and its group members for support and for clones and resources (<http://www.rzpd.de/>). We also thank J. Aaltonen, J. Buard, N. Creau, J. Gröet, R. Orti, J. Korenberg, M.C. Potier and G. Roizes for bacterial clones; D. Cox for discussions; A. Fortna, H.S. Scott, D. Slavov and G. Vacano for contributions; and N. Weizenbaum for editorial assistance. The RIKEN group is mainly supported by a Special Fund for the Human Genome Sequencing Project from the Science and Technology Agency (STA) Japan, and also by a Fund for Human Genome Sequencing from the Japan Society and Technology Corporation (JST) and a Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sport and Culture, Japan. The Jena group was supported by the Federal German Ministry of Education, Research and Technology (BMBF) through Projekträger DLR, in the framework of the German Human Genome Project, and by the Ministry of Science, Research and Art of the Freestate of Thuringia (TMWFK). The Keio group was supported in part by the Fund for Human Genome Sequencing Project from the JST, Grants-in-Aid for Scientific Research, and the Fund for "Research for the Future" Program from the Japan Society for the Promotion of Science (JSPS); they also received support from Grants-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports and Culture of Japan. The Braunschweig group was supported by BMBF through Projekträger DLR, in the framework of the German Human Genome Project. The MPIMG-Berlin group acknowledge grants from BMBF through Projekträger DLR in the framework of the German Human Genome Project and from the EU. Support also came from the Boettcher Foundation, NIH, Swiss National Science Foundation, EU and MRC. We also thank E. Wain and the Human Gene Nomenclature Committee for working out the chromosome 21 gene symbols, and Y. Groner for cloning and sequencing the first gene on chromosome 21 (SOD).

Correspondence and requests for materials should be addressed to Y.S. (e-mail: sakaki@gsc.riken.go.jp), A.R. (e-mail: andrex1x@aol.com), N.S. (e-mail: shimizu@dmf.med.keio.ac.jp), H.B. (e-mail: bloecker@gbf.de) or M.L.Y. (e-mail: yasp@molgen.mpg.de). Genomic clones can be requested from any of the five groups. Detailed clone information, maps, FISH data, annotated gene catalogue, gene name alias and supporting data sets are available from the RIKEN and MPIMG web sites (see Methods). Interactive chromosome 21 databases (HSA21DB) are maintained at MPIMG and RIKEN. All sequence data can be obtained from Genbank, EMBL and DDBJ. They are also available from the individual web pages.

Discovery of a comet by its Lyman- α emission

J. Teemu T. Mäkinen*, Jean-Loup Bertaux†, Harri Laakso*, Tuija Pulkkinen*, Tuula Summanen*, Erkki Kyrölä*, Walter Schmidt*, Eric Quémerais† & Rosine Lallement†

* Finnish Meteorological Institute, Geophysical Research, PO Box 503, FIN-00101 Helsinki, Finland

† Service d'Aéronomie, BP 3, F-91371 Verrières-le-Buisson, France

Several searches for near-Earth objects have recently been initiated, as a result of increased awareness of the hazard of impacts on the Earth. These programs mainly search for asteroids, so amateur astronomers can still contribute to the discovery of comets, especially out of the orbital plane of the Solar System. An ideal way to search for comets would be to use a spaceborne instrument capable of imaging the whole sky on a daily basis in a systematic and repeatable way. Such an instrument already exists on the solar observatory SOHO; it operates at the Lyman- α wavelength of neutral hydrogen, which is the main component of the emission cloud of a comet. Here we report the discovery, using archival data from this satellite, of a hitherto unnoticed comet which reached a perihelion of 1.546 a.u. on 26 June 1997. We derive the water production rate of the comet as a function of time and find that it increases after perihelion, like that of comet Halley.

The primary scientific objective of the SWAN instrument¹ is to map the latitudinal distribution of solar wind deducible from

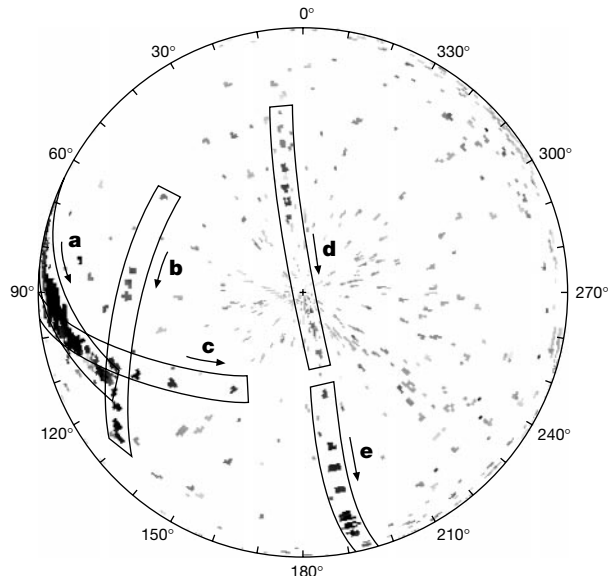


Figure 1 Bright comets visible in the southern ecliptic hemisphere from May to July 1997. The full-sky images were processed in a cartesian ecliptic projection with 1° resolution. We eliminated the large-scale interplanetary background by subtracting a 5° × 5° median filtered version of the image. Three months of observations at a time were fed to a neural network with one node for each pixel. The nodes were configured to excite or suppress their neighbours to indicate areas where, in consecutive images, changes characteristic of an object moving with an approximately constant velocity took place. **a–e**, The image displays the firing intensity of the nodes. Arrows depict the movement directions of comets Hale–Bopp (**a**), C/1997 N1 Tabur (**b**), 2P/Encke (**c**), K2 (**d**) and C/1997 O1 Tilbrook (**e**). The depicted five comets are all clear identifications, as indicated by the firing intensity. Several other trails can be seen with less confidence but it is difficult to distinguish between actual comets and instrumental effects.

asymmetries in the cavity that it carves in the passing cloud of interstellar neutral hydrogen, which resonantly scatters solar Lyman- α light. The use of SWAN full-sky images in cometary studies is impeded by the 1° × 1° resolution of the instrument and the 13-second exposure time in any particular direction. This limits the brightness of comets discernible on SWAN images, corresponding roughly to a visual magnitude of 12. The 1° pointing inaccuracy of the instrument smears light from bright ultraviolet stars, which are particularly abundant in the galactic plane and effectively render a large part of the sky unsuitable for cometary studies. On the other hand, the orbit of SOHO around the first Lagrange point between the Earth and the Sun enables observations which are devoid of the Earth's exospheric Lyman- α emission. Between the launch of SOHO in December 1995 and its loss in July 1998, SWAN produced 309 maps of varying sky coverage. Fortunately, SOHO was recovered some months later and continues to yield valuable data.

To find all comets that were in the observing range of the instrument, we processed the entire set of sky images up to June 1998 to single out potential comets. The aforementioned limitations make simple methods of detecting cometary objects insufficient, so we designed a combination of digital filtering and a neural network procedure. With correct parametrization, the method proved to be very effective, finding comets to above the noise limit. The detected comets could be identified by comparing their trails on the node map to the ephemerides of known comets. In this way we found 18 comets from the images, which naturally included such bright comets as C/1995 O1 Hale–Bopp, C/1996 B2 Hyakutake and C/1998 J1 SOHO but also less pronounced ones such as 46P/Wirtanen, which is the target of the forthcoming European Space Agency Rosetta mission. From May to July 1997, five comets were visible on the southern ecliptic hemisphere (Fig. 1), only four of which were previously known. The one passing through the southern ecliptic

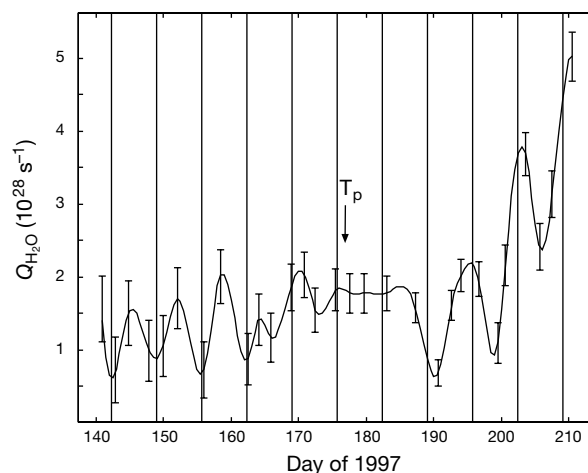


Figure 2 Water production rate Q_{H_2O} of comet K2 as a function of time. When determining Q_{H_2O} , the background contribution was eliminated by subtracting a comet observation from another one about one week apart, which left a combination of positive and negative images of the comet at different locations. A simple Haser model^{9,10} was then fitted to the combined image with position and Q_{H_2O} of the comet at both instances as free parameters. The statistical error estimation was based on the χ^2 values given by the Singular Value Decomposition procedure. The combined effect created by changes in solar activity and instrument calibration were estimated by observing the intensity of Lyman- α back-scattering from an area devoid of stars on the southern ecliptic hemisphere. During the visibility period of the comet the change in measured intensity is nearly monotonic and thus cannot account for the observed characteristics of Q_{H_2O} . That and model limitations add up to 25% of systematic error to the results. The thin line shows a spline interpolation between values. Compared to the preperihelion period denoted by vertical lines, the postperihelion activity appears to be shifted by about half a cycle. The distance of the comet from the sun during this period was almost constant at 1.6 ± 0.05 a.u. whereas the distance between the comet and SOHO decreased from over 1.8 a.u. to 1.17 a.u. and increased back to 1.4 a.u. The comet passed its perihelion on 26 June 1997.

pole had been missed by all observers until it was identified from the SWAN images². The orbital calculations revealed the comet now known as C/1997 K2 to be a very long-periodic comet with a perihelion distance of 1.546 a.u.

The curve of derived water production rate $Q_{\text{H}_2\text{O}}$ for comet K2 (Fig. 2) is rather flat near the perihelion because the heliocentric distance of the comet does not change much during the interval. At the perihelion, $Q_{\text{H}_2\text{O}}$ is around $1.8 \times 10^{28} \text{ s}^{-1}$, or about 500 kg s^{-1} of released water, which is 100 times less than that of comet Hale-Bopp at the same distance³. The $Q_{\text{H}_2\text{O}}$ of comet 1P/Halley was an order of magnitude higher⁴. Assuming an r^{-a} variation in the outgassing activity with heliocentric distance (where $a = 3-4$ for a very long-periodic comet), we can estimate that $Q_{\text{H}_2\text{O}}$ of comet K2 at 1 a.u. would have been of the order of 10^{29} s^{-1} . The $Q_{\text{H}_2\text{O}}$ curve also displays periodic fluctuations of the order of six days, indicating that the rotational period may be similar.

On the basis of extensive observations of comet Halley, we know that deriving the rotation rate of the nucleus from light-curve observations can be complex⁵. Here the rotational period is constant before perihelion, and after perihelion the same rotation appears with a phase shift of three days. Because the relative direction of sunlight changes by about 45° during the observation period, this supports the hypothesis that comet K2 has only one active source region and the six-day fluctuation is caused by a rotation in the orbital plane of the comet. After perihelion, $Q_{\text{H}_2\text{O}}$ increases significantly, but then the comet is lost among bright stars of the galactic plane. This kind of postperihelion activity was also observed on comet Halley, whereas many new comets, like comets Austin and Wilson, show decreasing outgassing activity towards perihelion and in the postperihelion region⁶. The viable mechanisms of cometary outbursts have been discussed widely⁷. In the case of comet K2 the six-day pattern indicates that this enhanced activity may arise from the original source region, which probably expanded during perihelion because of the heating of the nucleus. On the basis of these results, we argue that the comet is not dynamically new, although the orbital period cannot be estimated from the data.

We are not aware of any visual observations of the comet but, on the basis of semiempirical studies on the relationship between the visual magnitude and water production of a comet⁸, we estimate that the comet reached a visual magnitude of 11. Because the comet was almost constant in brightness over several months, it should have been easily observable from the ground. Another comet with high inclination, C/1997 O1 Tilbrook, passed the same region just one month before but was not discovered until near the ecliptic plane, which underlines the need for full-sky surveillance of comets. □

Received 4 February; accepted 14 March 2000.

- Bertaux, J.-L. *et al.* SWAN: a study of solar wind anisotropies on Soho with Lyman- α sky mapping. *Sol. Phys.* **162**, 403–439 (1995).
- Mäkinen, T. & Bertaux, J.-L. *IAU Circ. No. 7327* (1999).
- Combi, M. R., Reinard, A. A., Bertaux, J.-L., Quémerais, E. & Mäkinen, T. SOHO/SWAN observations of the structure and evolution of the hydrogen Lyman- α coma of comet Hale-Bopp (1995 O1). *Icarus* **144**, 191–202 (2000).
- Feldman, P. D. in *Comets in the Post-Halley Era Vol. 1* (eds Newburn, R. L. Jr *et al.*) 139–148 (Kluwer Academic, Dordrecht, The Netherlands, 1991).
- Festou, M. C. *et al.* Periodicities in the light curve of P/Halley and the rotation of its nucleus. *Astron. Astrophys.* **187**, 575–580 (1987).
- Meech, K. J. in *Comets in the Post-Halley Era Vol. 1* (eds Newburn, R. L. Jr *et al.*) 629–669 (Kluwer Academic, Dordrecht, The Netherlands, 1991).
- Hughes, D. W. in *Comets in the Post-Halley Era Vol. 2* (eds Newburn, R. L. Jr *et al.*) 825–851 (Kluwer Academic, Dordrecht, The Netherlands, 1991).
- Roettger, E. E., Feldman, P. D., A'Hearn, M. F. & Festou, M. C. Comparison of water production rates from UV spectroscopy and visual magnitudes for some recent comets. *Icarus* **86**, 100–114 (1990).
- Bertaux, J.-L. *et al.* Lyman- α observations of comet Hyakutake with SWAN on SOHO. *Planet. Space Sci.* **46**, 555–568 (1998).
- Festou, M. C. The density distribution of neutral compounds in cometary atmospheres. I. Models and equations. *Astron. Astrophys.* **95**, 69–79 (1981).

Acknowledgements

SOHO is an international cooperative mission of ESA and NASA. SWAN was financed in

France by CNES with support from CNRS and in Finland by TEKES and the Finnish Meteorological Institute. The work of J.T.T.M., H.L. and T.S. was supported by the Academy of Finland.

Correspondence and requests for materials should be addressed to J.T.T.M. (e-mail: Teemu.Makinen@fmi.fi).

Thermally fluctuating superconductors in two dimensions

Subir Sachdev & Oleg A. Starykh

Department of Physics, Yale University, PO Box 208120, New Haven, Connecticut 06520-8120, USA

In many two-dimensional superconducting systems^{1–4}, such as Josephson-junction arrays, granular superconducting films, and the high-temperature superconductors, it appears that the electrons bind into Cooper pairs below a pairing temperature (T_p) that is well above the Kosterlitz–Thouless temperature (T_{KT} , the temperature below which there is long-range superconducting order^{5–10}). The electron dynamics at temperatures between T_{KT} and T_p involve a complex interplay of thermal and quantum fluctuations, for which no quantitative theory exists. Here we report numerical results for this region, by exploiting its proximity to a $T = 0$ superconductor–insulator quantum phase transition^{11,12}. This quantum critical point need not be experimentally accessible for our results to apply. We characterize the static, thermodynamic properties by a single dimensionless parameter, $\gamma(T)$. Quantitative and universal results are obtained for the frequency dependence of the conductivity, which are dependent only upon $\gamma(T)$ and fundamental constants of nature.

For clarity, we will present our results in the context of a familiar microscopic model for the superconductor–insulator transition; however, our results generalize to a much wider class of systems, and this will be discussed towards the end of this Letter. We consider an array of superconducting quantum dots at the sites, i , of a regular two-dimensional lattice. The operator $\hat{n}_i$ measures the number of electron pairs on dot i , and $\hat{\phi}_i$ is its canonically conjugate phase ($[\hat{n}_i, \hat{\phi}_j] = \delta_{ij}$). We consider the hamiltonian¹³

$$H = (E_C/2) \sum_i (\hat{n}_i - N_0)^2 - E_J \sum_{\langle ij \rangle} \cos(\hat{\phi}_i - \hat{\phi}_j) \quad (1)$$

where N_0 is the mean number of Cooper pairs on each dot and is required to be an integer, $\langle ij \rangle$ represents nearest neighbour pairs, E_C is the charging energy of a dot, and E_J is the Josephson tunnel coupling between dots. This model exhibits a superconductor to insulator transition as the dimensionless parameter $g = E_C/E_J$ is increased through a critical value g_c . The phase diagram in the T – g plane is summarized in Fig. 1. We will be mainly concerned with the region $g \leq g_c$. For $T < T_{KT}$, in the superconducting phase, the conductivity has the delta-function contribution $\text{Re}[\sigma] = (e^* \pi \rho_s(T)/\hbar^2) \delta(\omega)$, where $e^* = 2e$. This defines the superfluid stiffness $\rho_s(T)$ (in units of energy). The hyperscaling property of the quantum critical point implies that all dynamic properties in its vicinity are entirely characterized by a single energy, $\rho_s(0)$, which vanishes at $g = g_c$. In particular, the conductivity obeys¹²

$$\sigma(\omega, T) = \frac{e^{*2}}{h} \Sigma \left(\frac{\hbar \omega}{k_B T}, \frac{\rho_s(0)}{k_B T} \right) \quad (2)$$

where Σ is a completely universal scaling function. Such a two-argument scaling form is not overly constraining, and permits a rich

variety of behaviour which we shall describe here as a function of the second argument $\lambda \equiv \rho_s(0)/k_B T$. Earlier studies of superfluid dynamics (made without reference to quantum phase transitions) emerge in limiting regimes, with equation (2) only placing restrictions on certain parameter values.

Demanding consistency of equation (2) with the definition of $\rho_s(T)$ above immediately leads to the following conclusions^{14,15}: (1) $\rho_s(T)/\rho_s(0)$ is a universal function of λ ; (2) the Kosterlitz–Thouless transition occurs at the universal value $\lambda = \lambda_c$ where Σ is singular; (3) $T_{KT} = \rho_s(0)/\lambda_c$. Aspects of the data on the high-temperature superconductors are consistent with such trends^{5,10}, suggesting the proximity of an insulator of Cooper pairs (with stripe order).

We now describe the evolution of the following dynamic properties with decreasing λ : (1) Phase hydrodynamics. At very large $\lambda(k_B T \ll \rho_s(0))$ vortex excitations are exponentially suppressed, and we can derive an effective quantum action, S_φ for the continuum phase variable $\varphi(x, \tau)$ (where x is a two-dimensional spatial coordinate and τ is imaginary time) in a gradient expansion; the resulting action is valid only at the longest scales, and is similar to the ‘chiral lagrangians’ of particle physics:

$$S_\varphi = \int d^2x d\tau \left[\frac{\rho_s(0)}{2} (\partial_\mu \varphi)^2 - \frac{A_1 (\hbar v)^2}{2\rho_s(0)} (\partial_\mu \varphi)^2 (\partial_\nu \varphi)^2 \right].$$

Here μ, ν are spacetime indices with $\partial_\mu = (\nabla_x, \partial_\tau/v)$ and v is the velocity of the ‘spin-wave’ excitations produced by the harmonic terms in S_φ (v remains non-singular through the quantum-critical point). The non-linear term arises from integrating out amplitude fluctuations and leads to spin-wave scattering; consistency of the resulting transport properties with equation (2) demands that the scattering cross-section be universally determined by $\rho_s(0)$ —hence the dimensionless number A_1 is *universal*. Determination of the $T > 0$ transport properties of S_φ requires solution of the appropriate quantum kinetic equations (not shown here). The procedure is closely analogous to early work^{16,17} on phonon transport in Galilean-invariant superfluids; an important difference is that these systems had a cubic non-linearity which is forbidden in our case by particle-hole symmetry.

(2) Vortex hydrodynamics. In the shaded region of Fig. 1, in the vicinity of T_{KT} , the vortices proliferate, and the dynamics can be

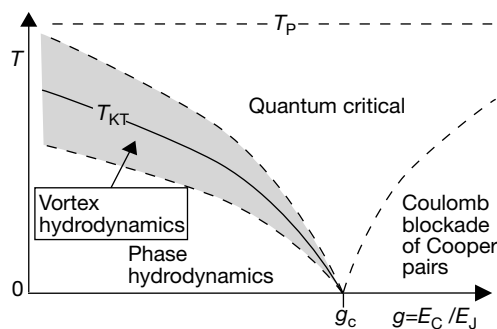


Figure 1 Phase diagram of H in equation (1). Dashed lines are crossover, while the solid line is a phase transition. There is superfluidity, with $\rho_s(T) > 0$, for $g < g_c$, $T < T_{KT}$; at $T = 0$, $\rho_s(0) \approx (g_c - g)^{2\nu}$ with ν the correlation length exponent ($\nu = 1$ for H). The ground state for $g > g_c$ is an insulator. If one studies H by a tunnelling probe to a good superconductor, there will be a zero-voltage Josephson current in the region where $\rho_s(T)$ is nonzero, while for $g > g_c$ and low T , a Cooper-pair current will only flow upon applying a finite voltage which overcomes the Coulomb blockade. The region of phase hydrodynamics is described by the action S_φ : in general spatial dimension d , the coefficient of the quartic term is $(A_1/2)(\hbar v)^{2(d-1)}[\rho_s(0)]^{(d-3)(d-1)}$, and the result for A_1 in an expansion¹⁴ in $\epsilon = 3 - d$ is $A_1 = [10(4\pi)^{-(d+1)/2}/(\Gamma((d+1)/2)\epsilon)]^{2/(d-1)} \times (1 - 11\epsilon/30 + O(\epsilon^2))$. In the quantum-critical region the conductivity obeys equation (9); at $g = g_c$, $\gamma(T)$ takes a T -independent value which can be computed²¹ in a $1/N$ expansion: $\gamma(T) = [5/(4\pi \ln((5 + 1/2)))^{1/2} (1 - 0.5468/N)]$.

described by a well-developed classical theory¹⁸ for the vortices alone. This theory is contained within Σ for $\lambda \approx \lambda_c$, and compatibility constrains various prefactors. For example, for $T > T_{KT}$ the response of the free vortices is controlled by their diffusivity, $A_2 \hbar v^2 / \rho_s(0)$, and their screening length, $A_3 (\hbar v / \rho_s(0)) e^{A_4 / (\lambda_c - \lambda)^{1/2}}$; in general $A_2 - A_4$ are arbitrary dimensionless scale factors—however, near $g = g_c$ they become universal numbers.

(3) Quantum critical. Discussion of this small λ regime will occupy the remainder of this work. Previous work^{12,19,20} relied upon expansions in either small $3 - d$, or small $d - 1$, or large N (the number of real order parameter components). Here we shall present a theory directly for the physical case $N = 2$, $d = 2$.

Unlike the two previous regimes discussed above, it is no longer possible to decouple the spin-wave and vortex degrees of freedom. We will therefore use the complex superconducting order parameter $\psi(x, t)$ (t is real time) which is the continuum limit of the lattice operator $e^{i\varphi}$, and allow for both amplitude and phase fluctuations in ψ . Our theory follows from two hypotheses: First, the equal-time correlations of ψ are controlled by a gaussian effective action. Evidence for this hypothesis emerged in detailed studies of order parameter correlations in the quantum-critical region²¹; the non-gaussian components of the ψ correlations are weak because the order parameter anomalous dimension at the quantum-critical point, $\eta \approx 0.03$, is so small. (This small η also shows why ψ is preferred over the ‘dual-order parameter’ measuring vorticity²²—the latter has an appreciable anomalous dimension²³.) The second hypothesis is that the time evolution of $\psi(x, t)$ is described by classical equations of motion. The characteristic relaxation (phase coherence) time in the quantum-critical regime is of order²⁴ $\hbar/k_B T$; however, the dominant spectral weight is at frequencies $\hbar\omega < k_B T$, and there is good evidence^{12,15} that the errors made by focusing on this low-frequency classical regime are quite small. We note that a different classical dynamic model was considered in ref. 15, but it does not apply to transport properties.

We now define our model for quantum-critical transport, and then present its exact (numerical) solution. In addition to the field $\psi(x, t)$, we will need the canonically conjugate variable measuring density fluctuations, $\delta n(x, t)$, which is the continuum limit of $\hat{n}_i - N_0$. The equal-time correlations of ψ and δn are described by the partition function

$$Z = \int D\psi(x) D\delta n(x) \exp(-H_1 + H_2/k_B T) \quad (3)$$

$$H_1 = \int d^2x [|\nabla\psi|^2 + \xi^{-2}(T)|\psi|^2] \quad (4)$$

$$H_2 = (1/[2\chi_u(T)]) \int d^2x (\delta n(x))^2 \quad (5)$$

The overall scale of ψ is arbitrary, and has been adjusted to obtain a unit coefficient for the gradient term in equation (4). The correlation length, $\xi(T)$, which is assumed to be larger than all microscopic scales, and the compressibility, $\chi_u(T)$, are determined by the underlying quantum physics, and will generally be regarded as unknown input parameters in our dynamic theory; proximity to the quantum-critical point does (weakly) restrict their allowed T dependencies, but we can envisage a more general application of our dynamic theory, without reference to quantum criticality, in which case $\xi(T)$, $\chi_u(T)$ are arbitrary. The time evolution of ψ , δn , is defined by a minimal model—the Hamilton–Jacobi equations of $H_1 + H_2$, and the only non-zero Poisson bracket

$$\{\delta n(x), \psi(x')\}_{P.B.} = (i/\hbar) \psi(x) \delta^2(x - x'), \quad (6)$$

which is the continuum, classical limit of the commutator between $\hat{\varphi}_i$ and $\hat{n}_i$. Notice that $\hbar$ appears on the right-hand side even though we are considering classical equations. The equations of motion implied by equations (4)–(6) are simple and familiar. They are the

continuity equation

$$\partial \delta n(x, t) / \partial t + \nabla \cdot J(x, t) = 0 \quad (7)$$

where $J = -(i/\hbar)(\psi^* \nabla \psi - \psi \nabla \psi^*)$, and

$$\partial \psi(x, t) / \partial t = -(i/\hbar) \Phi(x, t) \psi(x, t) \quad (8)$$

which is the Josephson equation, with the electrochemical potential $\Phi(x, t) = \delta n(x, t) / \chi_u(T)$. We re-state our dynamical model for quantum critical transport: choose a set of equal-time initial conditions for ψ and δn from the thermal ensemble defined by Z , and evolve them deterministically using equations (7) and (8). Correlation functions are determined by the average over initial conditions.

We will shortly present strong numerical evidence that equations (3)–(8) define a sensible continuum theory free of short-distance or short-time ('ultraviolet') divergences; this is supported by perturbative and renormalization group arguments, which we do not describe here. Unlike S_ϕ , the present theory describes the couplings of vortex and spin-wave fluctuations at different length scales. It is helpful to visualize the continuum theory by coarse-graining to a lattice spacing, a : the shorter-distance degrees of freedom lead to fluctuations in the phase and amplitude of ψ , but the long-distance transport properties are insensitive to the value of a . The absence of factors of a in the final results allows us to deduce their functional dependence on $\xi(T)$, $\chi_u(T)$ by simple engineering dimensional analysis. In this manner we conclude that

$$\sigma(\omega, T) = \frac{e^*{}^2}{h} \gamma(T) \Sigma_c \left(\gamma(T) \frac{\hbar \omega}{k_B T} \right) \quad (9)$$

where Σ_c is a universal function ('exact' numerical results are below), and the dimensionless $\gamma(T)$ is defined by

$$\gamma(T) \equiv [k_B T \chi_u(T) \xi^2(T)]^{1/2} \quad (10)$$

Consistency of equation (9) with equation (2) only requires that

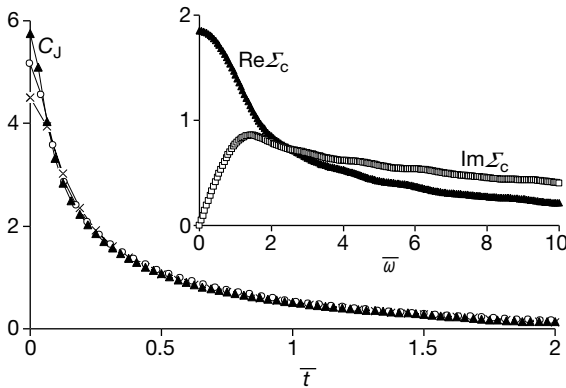


Figure 2 Results of our numerical simulation of equations (3)–(8). The lattice form of $H_1 + H_2$ was obtained by mapping the momentum (k_x, k_y) dependence of the couplings of the continuum hamiltonian under $k_x^2 \rightarrow (4/a^2)[K_1 \sin^2(k_x a/2) + K_2 \sin^2(k_x a) + K_3 \sin^2(3k_x a/2)]$, and similarly for k_y (for $|\nabla \psi|^2$, this amounts to including first, second and third neighbour couplings between the ψ). We chose two different sets of values for $K_{1,2,3}$ to test the independence of the continuum theory on the lattice realization—the familiar $K_1 = 1$, $K_2 = K_3 = 0$, and $K_1 = 3/2$, $K_2 = -3/20$, $K_3 = 1/90$ for which $k_x^2 \rightarrow k_x^2(1 + O(k_x a)^6)$. The initial conditions specified by Z form a gaussian ensemble, and are easily generated in a single sweep. The time evolution was carried out by a fourth order predictor-corrector algorithm with a time step determined to conserve total energy to a relative accuracy better than 10^{-5} . Shown is the current autocorrelation function, C_J , obtained by averaging 3,000 sweeps with the first lattice realization A values, $L = 64$, and $a/\xi(T) = 1/8$ (crosses), $\sqrt{2}/16$ (open circles), and $1/16$ (triangles). Inset, real and imaginary parts of Σ_c , obtained from C_J at $a/\xi(T) = 1/16$ by the Fourier transform $\Sigma_c(\omega) = \int_0^\infty dt C_J(t) e^{i\omega t}$; here $\bar{\omega} = \gamma(T) \hbar \omega / k_B T$, $\bar{t} = k_B T t / (\hbar \gamma(T))$. For large $|\bar{\omega}|$ we have $\text{Re} \Sigma_c \approx 1/|\bar{\omega}|$ and $\text{Im} \Sigma_c \approx \ln(|\bar{\omega}|)/\bar{\omega}$.

$\gamma(T)$ is a universal function of λ (also recall that the present theory is valid only for small λ). Exactly at $g = g_c$ ($\lambda = 0$), hyperscaling arguments imply that $\xi(T) \approx T^{-1/z}$ and $\chi_u(T) \approx T^{(2-z)/z}$; thus $\gamma(T)$ is a T -independent universal number for any z . The result, equation (9), has a clear experimental signature: it implies a correlation between the d.c. conductivity and the inverse frequency-width of $\sigma(\omega)$, as they are both determined by $\gamma(T)$. More stringent comparisons can be made by using the measured d.c. conductivity to determine the unknown $\gamma(T)$, and then using our numerical results below for Σ_c to determine the ω dependence of σ . It should be noted that our present computations for Σ_c will not apply for very large $\hbar \omega / k_B T$, as then a full quantum theory is necessary, and we cross over to the phase-coherent regime discussed earlier^{7,19}.

We turn to our numerical results for Σ_c . The simulations were carried out on an $L \times L$ square lattice of spacing a with periodic boundary conditions, and $a \ll \xi(T) \ll La$. For each a , successively larger values of L were used, until the results became L -independent, and the continuum limit was then approached as $a \rightarrow 0$. Our results are shown in Fig. 2. There is some a dependence in the current autocorrelation at small times (this is expected, and the results are consistent with a logarithmic dependence on a which can be computed theoretically). However, the a dependence quickly disappears at slightly later times, and then a universal continuum value obtains; this is strong evidence for the existence of the continuum theory of equations (3)–(8). We obtained $\Sigma_c(0) \approx 1.85$; for the model H in equation (1), this combines with the value of $\gamma(T)$ quoted in Fig. 1 to yield $\sigma(0, T) = 0.82(e^*{}^2/h)$ at $g = g_c$.

The experiments of Rimberg *et al.*¹ appear to be a convenient testing ground for our theory. The presence of the metallic gate leads to new terms, not contained in equation (1), in the effective hamiltonian, and these were discussed by Wagenblast *et al.*²⁵. The universality class of the quantum critical point could be modified^{30,26} and this would be manifested in changes in the static parameters $\xi(T)$, $\chi_u(T)$, and $\gamma(T)$. However, the metallic gate also has beneficial effects: first, it screens the long-range Coulomb interaction, and so justifies the short-range coupling in equation (5), and second, external dissipation reduces quantum interference effects, and justifies applicability of a classical dynamical model, beyond the arguments already given in refs 12 and 15. The effects of external dissipation on classical, collective, $T > 0$ dynamic models can be deduced from rather general arguments^{27–29} (unlike the situation for $T = 0$ quantum dynamics that we are not focusing on here). These imply that, for example, equation (8) is modified to $\partial \psi / \partial t = -\Gamma \delta H_1 / \delta \psi^* - (i/\hbar) \Phi \psi + \zeta$, where Γ is a damping coefficient and ζ is the associated random noise. However, it can be argued, as in refs 27–29, that for large $\xi(T)$, such modifications leave the dynamic function Σ_c unchanged. To summarize: external dissipation can be a relevant perturbation on the quantum critical point (and thus modify $\gamma(T)$), but it is an irrelevant correction to our effective classical model for $T > 0$ dynamics, and so equation (9), and our quantitative results for Σ_c in Fig. 2, apply unchanged to the system in ref. 1.

Next, we consider the inclusion of long-range Coulomb interactions. This will modify H_2 to $e^*{}^2 \int d^2 x d^2 x' \delta n(x) \delta n(x') / 2|x - x'|$, and leave equations (7) and (8) unchanged, but with $\Phi(x, t) = e^*{}^2 \int d^2 x' \delta n(x', t) / |x - x'|$. Assuming the existence of the continuum limit of the classical dynamic theory, we again obtain equation (9), but with $\gamma(T) = [k_B T \xi(T) e^*{}^2]^{1/2}$; the functional form of Σ_c will of course be changed and requires a separate numerical simulation. Note that at $g = g_c$, $\gamma(T)$ is a T -independent universal number only if $z = 1$, the value expected for long-range interactions⁷.

The experiments of Corson *et al.*⁴ on an underdoped high-temperature superconductor observe scaling closely related to equation (9), with the frequency scale proportional to the resistivity scale, and a scaling function with large ω behaviour similar to that in Fig. 2. Although the dominant T dependencies, and the structure of their observed dynamic scaling function agrees with equation (9),

their fits use a prefactor T_0^2 whose weak T -dependence disagrees with our continuum two-dimensional theory. We speculate that this can be explained by corrections to scaling: there is an appreciable density of fermionic carriers not bound into Cooper pairs, and this would reduce the effective bosonic superfluid density which enters our scaling forms with a T -dependent factor. Corson *et al.*⁴ motivated the scaling using the vortex theory¹⁸, but did not consider bound vortex pairs, whose contributions do not obey their scaling assumptions.

We have delineated the distinct dynamic regimes of thermally fluctuating superconductors: the low- T phase-only hydrodynamics, the classical vortex hydrodynamics in the vicinity of T_{KT} , and the quantum-critical region where phase and vortex fluctuations strongly coupled—here we proposed a dynamical model in which non-linear ‘mode-coupling’ terms demanded by the Poisson bracket, equation (6), dominate the universal, low-frequency, dissipative dynamics. Our approach could be extended to other two-dimensional systems, such as the magnetic-field-tuned superconductor–insulator transition, where a coupling to the external field would be required in equation (4), and quantum Hall transitions. □

Received 12 January; accepted 4 April 2000.

1. Rimberg, A. J. *et al.* Dissipation-driven superconductor–insulator transition in a two-dimensional Josephson-junction array. *Phys. Rev. Lett.* **78**, 2632–2635 (1997).
2. Chervenak, J. A. & Valles, J. M. Observation of critical amplitude fluctuations near the two-dimensional superconductor–insulator transition. *Phys. Rev. B* **59**, 11209–11212 (1999).
3. Marković, N., Christiansen, C., Mack, A. M., Huber, W. H. & Goldman, A. M. Superconductor–insulator transition in two dimensions. *Phys. Rev. B* **60**, 4320–4328 (1999).
4. Corson, J., Mallozzi, R., Orenstein, J., Eckstein, J. N. & Bozovic, I. Vanishing of phase coherence in underdoped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. *Nature* **398**, 221–223 (1999).
5. Uemura, Y. J. *et al.* Universal correlations between T_c and n/m (carrier density over effective mass) in high- T_c cuprate superconductors. *Phys. Rev. Lett.* **62**, 2317–2320 (1989).
6. Doniach, S. & Inui, M. Long-range Coulomb interactions and the onset of superconductivity in the high- T_c materials. *Phys. Rev. B* **41**, 6668–6678 (1990).
7. Fisher, M. P. A., Grinstein, G. & Girvin, S. M. Presence of quantum diffusion in two dimensions: Universal resistance at the superconductor–insulator transition. *Phys. Rev. Lett.* **64**, 587–590 (1990).
8. Trivedi, N. & Randeria, M. Deviations from Fermi-liquid behavior above T_c in 2D short coherence length superconductors. *Phys. Rev. Lett.* **75**, 312–315 (1995).
9. Emery, V. J. & Kivelson, S. A. Importance of phase fluctuations in superconductors with small superfluid density. *Nature* **374**, 434–437 (1995).
10. Emery, V. J. & Kivelson, S. A. Crossover and phase coherence in cuprate superconductors. *J. Phys. Chem. Solids* **59**, 1705–1710 (1998).
11. Sondhi, S. L., Girvin, S. M., Carini, J. P. & Shahar, D. Continuous quantum phase transitions. *Rev. Mod. Phys.* **69**, 315–333 (1997).
12. Sachdev, S. *Quantum Phase Transitions* (Cambridge Univ. Press, Cambridge, 1999).
13. Doniach, S. Quantum fluctuations in two dimensional superconductors. *Phys. Rev. B* **24**, 5063–5070 (1981).
14. Sachdev, S. Theory of finite-temperature crossovers near quantum critical points close to, or above, their upper-critical dimension. *Phys. Rev. B* **55**, 142–163 (1997).
15. Sachdev, S. Universal relaxation dynamics near two-dimensional quantum critical points. *Phys. Rev. B* **59**, 14054–14073 (1999).
16. Khalatnikov, I. M. *An Introduction to the Theory of Superfluidity* (W. A. Benjamin, New York, 1965).
17. Popov, V. N. *Functional Integrals in Quantum Field Theory and Statistical Physics* (D. Reidel, Boston, 1983).
18. Ambegaokar, V., Halperin, B. I., Nelson, D. R. & Siggia, E. D. Dissipation in two dimensional superfluids. *Phys. Rev. Lett.* **40**, 783–786 (1978).
19. Damle, K. & Sachdev, S. Nonzero-temperature transport near quantum critical points. *Phys. Rev. B* **56**, 8714–8733 (1997).
20. Herbut, I. F. Finite temperature transport at the superconductor–insulator transition in disordered systems. *Phys. Rev. Lett.* **81**, 3916–3919 (1998).
21. Chubukov, A. V., Sachdev, S. & Ye, J. Theory of two-dimensional quantum Heisenberg antiferromagnets with a nearly critical ground state. *Phys. Rev. B* **49**, 11919–11961 (1994).
22. Fisher, M. P. A. Quantum phase transitions in disordered two-dimensional superconductors. *Phys. Rev. Lett.* **65**, 923–926 (1990).
23. Sudbo, A., Nguyen, A. K. & Hove, J. Anomalous scaling dimensions and critical points in type-II superconductors. Pre-print cond.mat/9907386 at (xxx.lanl.gov) (1999).
24. Sachdev, S. & Ye, J. Universal quantum-critical dynamics of two-dimensional antiferromagnets. *Phys. Rev. Lett.* **69**, 2411–2414 (1992).
25. Wagenblast, K.-H., van Otterlo, A., Schön, G. & Zimányi, G. T. Superconductor–insulator transition in a tunable dissipative environment. *Phys. Rev. Lett.* **79**, 2730–2733 (1997).
26. Dalidovich, D. & Phillips, P. Fluctuation conductivity in insulator–superconductor transitions with dissipation. *Phys. Rev. Lett.* **84**, 737–740 (2000).
27. Kawasaki, K. Kinetic equations and time correlation functions of critical fluctuations. *Ann. Phys. (N.Y.)* **61**, 1–56 (1970).
28. Hohenberg, P. C., Siggia, E. D. & Halperin, B. I. Density-correlation function of liquid helium near T_λ in the symmetric planar model. *Phys. Rev. B* **14**, 2865–2874 (1976).
29. Grepmpel, D. Comment on “Low-temperature behavior of two-dimensional quantum anti-ferromagnets”. *Phys. Rev. Lett.* **61**, 1041 (1988).
30. Wagenblast, K.-H., van Otterlo, A., Schön, G. & Zimányi, G. T. New universality class at the superconductor–insulator transition. *Phys. Rev. Lett.* **78**, 1779–1782 (1997).

Acknowledgements

We thank J. Orenstein, D. Grepmpel and A. Sudbo for useful discussions, and the Division of Materials Theory of the National Science Foundation for support.

Correspondence should be addressed to S.S. (e-mail: subir.sachdev@yale.edu).

Optical microscopy using a single-molecule light source

J. Michaelis, C. Hettich, J. Mlynek & V. Sandoghdar

Fachbereich Physik und Optik-Zentrum Konstanz, Universität Konstanz, Fach M696, 78457 Konstanz, Germany

Rapid progress in science on nanoscopic scales has promoted increasing interest in techniques of ultrahigh-resolution optical microscopy. The diffraction limit can be surpassed by illuminating an object in the near field through a sub-wavelength aperture at the end of a sharp metallic probe^{1,2}. Proposed modifications^{3,4} of this technique involve replacing the physical aperture by a nanoscopic active light source. Advances in the spatial⁵ and spectral⁶ detection of individual fluorescent molecules, using near-field and far-field methods⁷, suggest the possibility of using a single molecule^{8,9} as the illumination source. Here we present optical images taken with a single molecule as a point-like source of illumination, by combining fluorescence excitation spectroscopy¹⁰ with shear-force microscopy¹¹. Our single-molecule probe has potential for achieving molecular resolution in optical microscopy; it should also facilitate controlled studies of nanometre-scale phenomena (such as resonant energy transfer) with improved lateral and axial spatial resolution.

Over nearly two decades the resolution in scanning near-field optical microscopy (SNOM) using aperture probes has levelled out at 3a value of ~ 50 nm. In addition to technological difficulties such as reproducibility of tip fabrication, the resolution in this mode of SNOM is fundamentally limited because the finite skin depth of real metals leads to a finite effective aperture size that one can achieve in the laboratory³. To overcome this problem, some researchers have pursued apertureless SNOM in which a sharp tip is used to scatter the near field of the sample¹², but in this configuration the resolution is limited by the radius of curvature of the probe used. The use of a nanoscopic active medium as a probe for SNOM has the advantage that its size can be reduced to that of a single atom or molecule. In this case, the resolution should only depend on the separation between the source and the sample.

Figure 1a shows the schematics of a combined scanning confocal microscope and scanning near-field optical microscope that constitutes the heart of our experimental setup¹³ operating at $T = 1.4$ K. A commercial piezo-electric scanner controls the position of the sample in front of a tapered optical-fibre probe. A microscope objective with a numerical aperture of 0.8 can be used to illuminate the sample or to collect the light from it. For adjusting the focus of the objective and for positioning the fibre probe, home-made piezo-driven translation stages are used¹⁴. We use the shear-force signal from a quartz tuning fork to monitor or to regulate the separation between the sample surface and the probe¹⁵. In the experiments described here, the end of the fibre probe contained a micron-sized *p*-terphenyl crystal (see Fig. 1b) that was doped with terrylene molecules at a nominal concentration of about 10^{-7} . To prepare such a probe, we first co-sublimated *p*-terphenyl and terrylene to obtain crystals of the order of a few micrometres in size on a cover slide. The spatial doping concentration of the

microcrystals was then examined by collective excitation of the terrylene molecules at $\lambda = 514$ nm in a confocal arrangement under ambient conditions^{16,17}. Once a microcrystal of a desired size and doping was located, it was carefully glued to the end of an uncoated chemically etched single-mode fibre, and this probe was transferred into the cryogenic setup indicated in Fig. 1a.

Individual terrylene molecules in the micro-crystal were identified by fluorescence excitation spectroscopy^{10,18} at $T = 1.4$ K. The light of a ring-dye laser operating at $\lambda \approx 578$ nm was first intensity stabilized using an electro-optical modulator and then coupled into the optical fibre leading to the micro-crystal in the cryostat. Because the sharp zero-phonon lines of the terrylene molecules are inhomogeneously distributed in frequency space, individual molecules can become excited when scanning the laser frequency. The red-shifted fluorescence of a selectively excited molecule at $\lambda \approx 630$ nm is then detected by an avalanche photodiode after passing through the microscope objective and spectral filters. In this manner, we could typically record resonance lines narrower than 50 MHz, count rates of 100 kHz and a signal-to-noise ratio of the order of 100:1 (Fig. 1c). Aside from the exquisitely narrow spectral lines and a high signal-to-noise ratio, our system also proves to be extremely photostable in contrast to the dye molecules that are typically studied at room temperature⁵. These features are the main reason for choosing this system for our present work.

Once the laser was tuned into resonance with a single terrylene molecule, the probe was moved nearer the sample until the onset of a shear-force signal was observed at a separation of about 20 nm. The sample was then retracted to a given fixed axial position from the tip and scanned laterally while monitoring the shear-force signal as a safety measure. The sample consisted of 25 nm high triangular aluminium islands arranged in a hexagonal lattice with a period of 1.7 μm on a cover slide¹⁹. Figure 2a displays a topography image of the region that was studied in this experiment recorded with an atomic force microscope after the sample was taken out of the cryostat. The arrow indicates an area where a metallic island is largely missing, which results in two nanostructures. The image in Fig. 2b shows the raw fluorescence signal of a single molecule

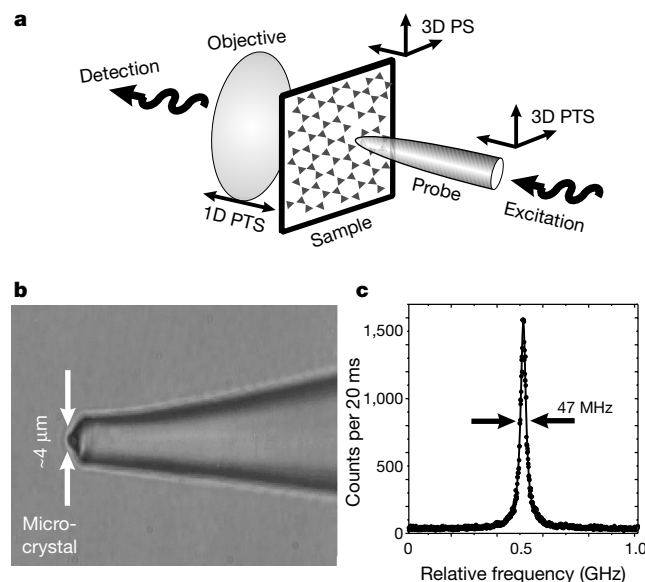


Figure 1 Overview of experimental features for realizing a single-molecule probe. **a**, Schematic view of the experimental configuration. PS, piezo-electric scanner; PTS, piezo-driven translation stage. See text for details. **b**, an optical microscope image of the fibre probe displaying a micro-crystal of *p*-terphenyl glued to its end. The micro-crystal was lightly doped with terrylene molecules. **c**, An excitation spectrum of the molecule used for recording the images in this experiment. The solid line shows a lorentzian fit to the data revealing a full-width at half-maximum of 47 MHz.

positioned in front of this sample just before the onset of the shear-force signal. The photon counts were integrated for 20 ms per pixel during which period the laser frequency was scanned back and forth once over an interval of ± 250 MHz around the resonance to compensate for small laser drifts and occasional spectral diffusions²⁰. As expected, the signal drops substantially each time a metallic island blocks the transmission of the molecular fluorescence. Although the triangular form of the islands does not manifest itself strongly, the image shows a striking reproducibility among different lattice periods. For example, the set of islands labelled *i* can clearly be distinguished from that labelled *ii* corresponding to the two possible orientations of the triangular structures. Also, the bright regions display a systematic pattern, with the exception of the area around the defect.

In Fig. 3 we show this image (Fig. 3e) together with three other optical images taken with the same single molecule at the same lateral position but with different distances between the crystal extremity and the sample (350 nm, 80 nm, 50 nm and 20 nm in Fig. 3a, b, d and e, respectively). The scans were recorded in the chronological order of Fig. 3b, a, e and d. These images are very lightly filtered by averaging the signal from each pixel with a weighting factor of 8 and its 8 neighbouring ones each with a weight of 1. To account for variations in the excitation intensity among different runs, the count rates in all four images have been normalized and a common colour scale is used so that one can compare them directly. Individual metallic islands and their hexagonal arrangement can be clearly identified in all cases, but the tiny defect structures give rise to a significant signal only at the smallest molecule-sample separation achieved in Fig. 3e. The fact that the contrast and the resolution are considerably worse at a larger distance also becomes evident when comparing Fig. 3a with the other images.

Bringing the sample closer in steps as small as 30 nm from Fig. 3b to d to e results in a substantial but continuous evolution of the bright and dark contrast patterns although these remain very robust within each figure. This can be seen more quantitatively in Fig. 3f where two cross-sections α and β exactly 1.7 μm apart are plotted. Note that the signal undergoes oscillations above and below the background indicated by a baseline. We attribute the origin of this effect and the rich images of individual metallic islands (for example, in Fig. 3e) to the polarization-sensitive interaction of the molecular dipole radiation with the aluminium triangles. Such interference-like phenomena have been reported in different

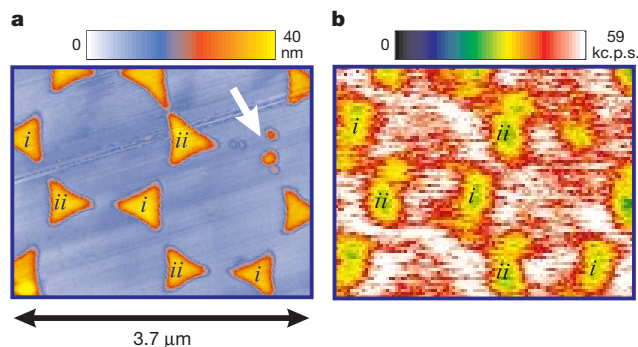


Figure 2 Topography and optical images of the sample. **a**, An atomic force microscope (AFM) topography image of the area of the sample used in this experiment. The arrow points to the remains of an island that is largely missing. We note that, although we have also recorded shear-force topography images of the sample taken at 1.4 K, because of our micrometre-size tip dimensions the resolution is not as good as that of an AFM. **b**, The optical raster-image recorded using the fluorescence of a single terrylene molecule for illumination. This scan contains 100×80 pixels with an integration time of 20 ms per pixel. The colour scale displays the raw experimental count rates in units of thousand counts per second (kc.p.s.).

experimental^{21,22} and theoretical^{23,24} works on conventional SNOM with polarized illumination. These studies have concluded that image interpretation becomes non-trivial for certain combinations of the electric field orientation and sample morphology and that the image quality depends on the illumination and detection polarization as well as the detection angle²². In the case of a single-molecule emitter, one should also take into account the modification of its spectral properties owing to the influence of the sample surface^{25,26}. A detailed theoretical and experimental study of these properties will be the topic of a future publication.

Clearly, the complex contrast observed in the images complicates the assignment of a resolution. When dealing with finite-sized structures, one common method for evaluating the resolution is measuring the edge sharpness of the obtained signal²⁷. In our case, however, following this guideline is not straightforward because of the oscillations about the background. We therefore merely note that the signal in Fig. 3f drops below the baseline from its 10% to 90% value over a length of about 180 nm which is less than one-third of the emission wavelength.

When extending the principles of scanning optical near-field microscopy to the case of a point-like electric dipole probe, the resolution is expected to depend only on the probe-sample separation. From a practical point of view, addressing molecules

that lie very close to the probe end should lead to higher spatial resolution and in the extreme limit to a molecular resolution. To achieve this, we plan to use sub-micrometre crystals and implement various techniques for identifying single molecules at the end of the crystal, for example, through evanescent excitation. In addition to the detection of the fluorescence intensity, the internal quantum mechanical properties of the molecule such as fluorescence lifetime^{26,28} and energy level shifts²⁶ can be used to extract information about its immediate environment. Our single-molecule probe could be also used to study a variety of phenomena such as surface-enhanced Raman spectroscopy and resonant energy transfer with an unprecedented lateral and axial spatial resolution. □

Received 20 January; accepted 15 March 2000.

- Lewis, A., Isaacson, M., Muray, A. & Harootunian, A. Scanning optical microscopy with 500 Å spatial resolution. *Biophys. J.* **41**, 405 (abstr.) (1983).
- Pohl, D. W., Denk, W. & Lanz, M. Optical stethoscopy: image recording with resolution $\lambda/20$. *Appl. Phys. Lett.* **44**, 651–653 (1984).
- Liebermann, K., Harush, S., Lewis, A. & Kopelman, R. A light source smaller than the optical wavelength. *Science* **247**, 59–61 (1990).
- Lewis, A. & Liebermann, K. Near-field optical imaging with a non-evanescently excited high-brightness light source of sub-wavelength dimensions. *Nature* **354**, 214–216 (1991).
- Betzig, E. & Chichester, R. J. Single molecules observed by near-field scanning optical microscopy. *Science* **262**, 1422–1425 (1993).
- Moerner, W. E. & Kador, L. Optical detection and spectroscopy of single molecules in a solid. *Phys. Rev. Lett.* **62**, 2535–2538 (1989).
- Basché, T., Moerner, W. E., Orrit, M. & Wild, U. P. *Single Molecule Optical Detection, Imaging and Spectroscopy* (VCH, Weinheim, Germany, 1997).
- Kopelman, R. & Tan, W. Near-field optics: imaging single molecules. *Science* **262**, 1382–1384 (1993).
- Sekatskii, S. K. & Letokhov, V. S. Single fluorescence centers on the tips of crystal needles: first observation and prospects for application in scanning one-atom fluorescence microscopy. *Appl. Phys. B* **63**, 525–530 (1996).
- Orrit, M. & Bernard, J. Single pentacene molecules detected by fluorescence excitation in a *p*-terphenyl crystal. *Phys. Rev. Lett.* **65**, 2716–2719 (1990).
- Betzig, E., Finn, P. L. & Weiner, J. S. Combined shear force and near-field scanning optical microscope. *Appl. Phys. Lett.* **60**, 2484–2486 (1992).
- Gross Levi, B. Progress made in near-field imaging with light from a sharp tip. *Phys. Today* **52**, 18–20 (1999).
- Michaelis, J. *et al.* A single molecule as a probe of the optical intensity distribution. *Opt. Lett.* **24**, 581–583 (1999).
- Bingelli, M. *et al.* Novel design for a compact fiber optic scanning force microscope. *Rev. Sci. Instrum.* **64**, 2888–2891 (1993).
- Karrai, K. & Grober, R. D. Piezoelectric tip-sample distance control for near-field optical microscopes. *Appl. Phys. Lett.* **66**, 1842–1844 (1995).
- Fleury, L., Sick, B., Zumhofen, G., Hecht, B. & Wild, U. P. High photo-stability of single molecules in an organic crystal at room temperature observed by scanning confocal microscopy. *Mol. Phys.* **95**, 1333–1338 (1998).
- Kulzer, F., Koberling, F., Christ, T., Mews, A. & Basché, T. Terrylene in *p*-terphenyl: single molecule experiments at room temperature. *Chem. Phys.* **247**, 23–34 (1999).
- Kummer, S., Basché, T. & Bräuchle, C. Terrylene in *p*-terphenyl: a novel single crystalline system for single molecule spectroscopy at low temperatures. *Chem. Phys. Lett.* **229**, 309–316 (1994).
- Fischer, U. & Zingsheim, H. P. Submicroscopic pattern replication with visible light. *J. Vac. Sci. Technol.* **19**, 881–885 (1981).
- Ambrose, P., Basché, T. & Moerner, W. E. Detection and spectroscopy of single pentacene molecules in a *p*-terphenyl crystal by means of fluorescence excitation. *J. Chem. Phys.* **95**, 7150–7163 (1991).
- Betzig, E., Trautman, J. K., Weiner, J. S., Harris, T. D. & Wolfe, R. Polarization contrast in near-field scanning optical microscopy. *Appl. Opt.* **31**, 4563–4568 (1992).
- Huser, T., Novotny, L., Lacoste, T., Eckert, R. & Heinzelmann, H. Observation and analysis of near-field optical diffraction. *J. Opt. Soc. Am. A* **16**, 141–148 (1999).
- Martin, O. J. F., Girard, C. & Dereux, A. Generalized field propagator for electromagnetic scattering and light confinement. *Phys. Rev. Lett.* **74**, 526–529 (1995).
- Martin, O. J. F. 3D simulations of the experimental signal measured in near-field optical microscopy. *J. Microscopy* **194**, 235–239 (1999).
- Novotny, L. Single molecule fluorescence in inhomogeneous environments. *Appl. Phys. Lett.* **69**, 3806–3808 (1996).
- Henkel, C. & Sandoghdar, V. Single molecule spectroscopy near structured dielectrics. *Opt. Comm.* **158**, 250–258 (1998).
- Harootunian, A., Betzig, E., Isaacson, M. & Lewis, A. Super-resolution fluorescence near-field scanning optical microscopy. *Appl. Phys. Lett.* **49**, 674–676 (1986).
- Barchiesi, D., Pagnot, T., Pieralli, C. & Van Labeke, D. Fluorescence scanning near-field microscopy (FSNM) by measuring the decay-time of a fluorescent particle. *Proc. SPIE* **2384**, 90–100 (1995).
- Kulzer, F., Kummer, S., Matzke, R., Bräuchle, C. & Basché, T. Single molecule optical switching in terrylene in *p*-terphenyl. *Nature* **387**, 688–691 (1997).

Acknowledgements

We are grateful to O. Martin and C. Henkel for many fruitful discussions. We thank B. Eiermann for contribution to the initial phase of the experiment, T. Kalkbrenner for the fabrication of the sample, and H. M. Ludwig for the AFM image in Fig. 2. We also thank P. Leiderer and his co-workers, and colleagues at Omicron Vakuum GmbH for help and

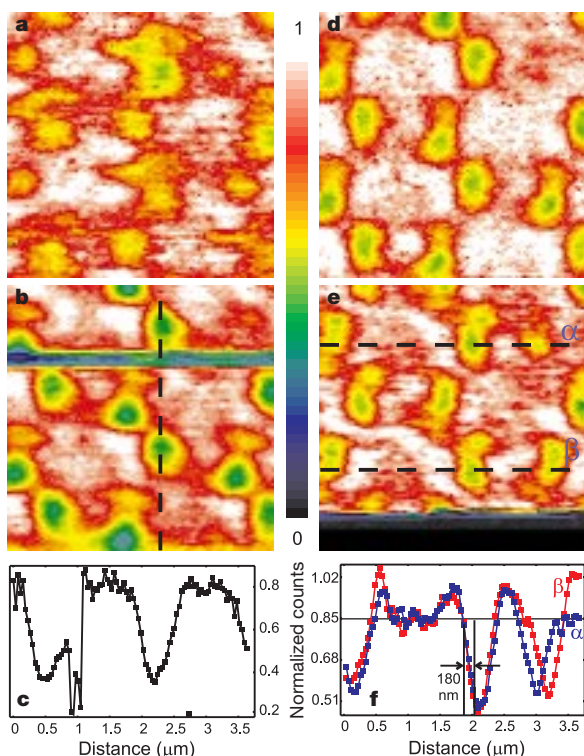


Figure 3 Images taken with the same molecule positioned at four different distances from the sample. These images are very lightly filtered (see text). The separation between the micro-crystal's extremity and the sample was 350 nm (**a**), 80 nm (**b**), 50 nm (**d**) and 20 nm (**e**). **e**, A more complete view of the run shown in Fig. 2b where in the last part the molecular spectrum became unstable and underwent a large jump of the order of hundred GHz (ref. 29) when the probe came into contact with the sample, so that a very small change in the shear-force signal was detected. **c**, The fluorescence signal along the cross-section shown in **b**. Note the sudden drop of the signal to ~200 counts per pixel (about 0.2 on the normalized scale) while our residual fluorescence corresponded to about 30 counts per pixel. Here the molecule has undergone a small spectral jump to the edge of the frequency scan range which was corrected for after a few scan lines by adjusting the laser frequency manually. **f**, the fluorescence signals along the two cross-sections α and β shown in **e**. A comparison of the four images taken over a period of over 2 h reveals the stability of the system and lack of any lateral drifts.

advice regarding the construction of the piezo-driven translation stages. J. M. acknowledges a fellowship from the Carl-Zeiss-Schott Förderstiftung. This work was supported by the Deutsche Forschungsgemeinschaft and the Bundesministerium für Bildung und Forschung.

Correspondence and requests for materials should be addressed to V.S. (e-mail: vahid.sandoghdar@uni-konstanz.de).

Intercalation of alkylamines into an organic polymer crystal

Akikazu Matsumoto*, Toru Odani*, Kazuki Sada†, Mikiji Miyata† & Kohji Tashiro‡

* Department of Applied Chemistry, Faculty of Engineering, Osaka City University, Sugimoto, Sumiyoshi-ku, Osaka 558-8585, Japan
† Graduate School of Engineering, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan
‡ Department of Macromolecular Science, Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan

Organic solid-state synthesis allows formation of products that are difficult or impossible to produce by conventional methods. This feature, and the high degree of reaction selectivity that can be achieved, is a direct result of the control over the relative orientation of the reactants afforded by the solid state. But as the successful development of 'topochemical reactions' requires the careful design of suitable reactant crystals, the range of both reactions and products amenable to this approach has been limited^{1,2}. However, recent advances in organic crystal engineering, particularly the rational design of complex solid architectures through supramolecular preorganization^{3–9}, have renewed interest in topochemical reactions. Previously, we have orientated muconate monomers—diene moieties with a carboxylate group on each end—using long-chain *n*-alkylammonium ions, such that the topochemical photopolymerization of the solid-state reactants produces layered crystals of stereoregular and high-molecular-mass polymers^{10–13}. Here we show that these polymer crystals are capable of repeated, reversible intercalation¹⁴ by conversion to the analogous poly(carboxylic acid), followed by transformation into a number of poly(alkylammonium

muconate)s upon addition of the appropriate amine. Introduction of functional groups into these crystals may allow the design of organic solids for applications such as molecular recognition, separation and catalysis, thereby extending the range and practical utility of current intercalation compounds^{15–20}.

Alkylammonium (*Z,Z*)-muconates undergo polymerization in the crystalline state under photoirradiation. When a higher *n*-alkyl group (carbon number $m \geq 10$) is introduced in the *N*-substituent, the photoproduct is an insoluble polymer crystal (Fig. 1; see Supplementary Information for preparation and polymerization procedures). Derivatives with a lower *n*-alkyl substituent ($m < 10$) favour topotactic isomerization to the corresponding (*E,E*)-isomers. In Table 1, the results of the photopolymerization are given together with the values of 2θ and interplanar distance (d) determined by small-angle powder X-ray diffraction for both monomer and polymer crystals. No peak was observed in the small-angle region for the monomer crystals with $m < 10$, in contrast to the intense peaks at $2-3^\circ$ for the monomer and polymer crystals with $m \geq 10$ (Fig. 2a). They correspond to 30–45 Å of the d value, which increased in proportion to the number of carbons in the *N*-alkyl group. The topochemically polymerizable (*Z,Z*)-muconates have similar crystal lattice parameters to each other^{11,12}: monoclinic, $P2_1/c$, $a = 10.2320$ Å, $b = 4.9310$ Å, $c = 11.4970$ Å, $\beta = 107.146^\circ$, $V = 554.29$ Å³, $Z = 2$ for the ethyl ester; monoclinic, $P2_1/a$, $a = 10.98$ Å, $b = 4.862$ Å, $c = 17.72$ Å, $\beta = 97.93^\circ$, $V = 936$ Å³, $Z = 4$ for the benzylammonium salt; and monoclinic, $P2_1/a$, $a = 11.033$ Å, $b = 4.9360$ Å, $c = 18.011$ Å, $\beta = 95.055^\circ$, $V = 977.0$ Å³, $Z = 4$ for the 2-chlorobenzylammonium salt. A lamellar structure of alkylammonium cation and muconate dianion layers is observed in the monomer crystals of the benzylammonium derivatives, and this ordered structure remains during the topochemical polymerization. Polymerization behaviour that depends on the length of the alkyl

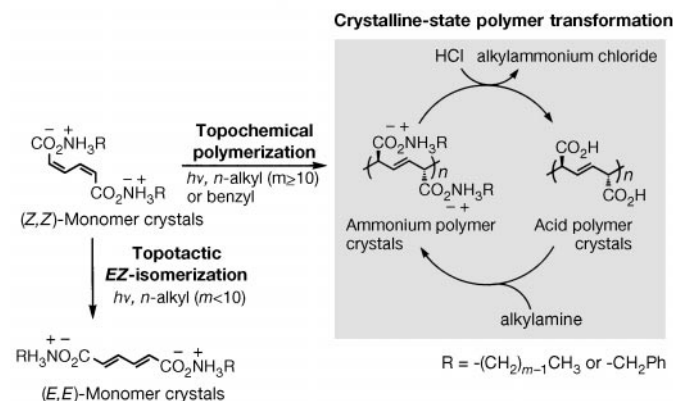


Figure 1 Photoreactions and polymer transformation performed in the crystalline state. Depending on the carbon number of the *n*-alkylammonium group, alkylammonium (*Z,Z*)-muconate undergoes either topochemical polymerization or topotactic *EZ*-isomerization when exposed to photoirradiation in the crystalline state. The ammonium polymer crystals obtained by the topochemical polymerization are converted in the crystalline state to acid polymer crystals via polymer transformation. The reverse reaction from polyacid to the ammonium polymers also occurs.

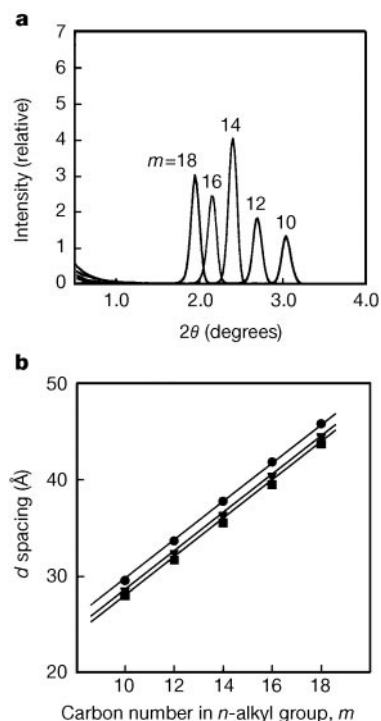


Figure 2 Small-angle X-ray scattering. **a**, Diffraction profiles of poly(*n*-alkylammonium (*Z,Z*)-muconate) crystals as polymerized. **b**, Relationship between the carbon number (m) of the *n*-alkyl group and the d values for *n*-alkylammonium (*Z,Z*)-muconates. Circles, monomer crystals; squares, polymer crystals obtained as polymerized under photoirradiation in the crystalline state after removing unreacted monomers; triangles, polymer crystals obtained by the crystalline-state polymer transformation from poly(muconic acid) with corresponding *n*-alkylamines.

chains in the *N*-substituents is because of the presence or absence of such a lamella structure in the crystals; that is, the longer alkyl chains favour the lamella-type molecular assembly appropriate for the topochemical polymerization in the crystals.

The polymer obtained can be transformed to other kinds of alkylammonium polymers via poly(muconic acid) by heterogeneous polymer reactions (Fig. 1). We dispersed the polymer crystals obtained by photopolymerization in acidic methanol at room temperature with stirring for 30 min to convert them to poly(muconic acid) crystals. The powder X-ray diffraction of the acid polymer crystals indicated a peak with $d = 4.8 \text{ \AA}$, but no peak was observed in the small-angle region (see Supplementary Information). The isolated acid polymer was reversely transformed into the ammonium polymers by reaction with the corresponding amine. We observed a strikingly visible change in crystal volume during the transformation, as expected. These transformations were completely heterogeneous because of the solubility of the polymer crystals both as the acid or ammonium derivatives. Nevertheless, the transformation to various *n*-alkylammonium polymers occurred with a high efficiency (94–96% conversion; Table 1). The d values agreed well with those for the untransformed crystals as polymerized. The repeated cycles of the ammonium–acid transformation also provided similar polymer crystals without the

collapse of the layer structure and change in the crystal habit (see Supplementary Information). When we used smaller alkylamines, the efficiency of the incorporation decreased, and the alkylammonium group did not have a well-ordered structure under similar conditions.

Figure 2b shows the comparison of the m dependence of the d values observed for the monomer crystals, the polymer crystals without transformation, and the polymers prepared by transformation via polymer reactions. The graphs for each of the crystals have similar gradients, indicating that the thickness of the alkylammonium layer increased by $1.0 \text{ \AA}$ for each carbon in the *N*-alkyl substituent. This distance increment suggests an alkyl-chain structure with a tilt angle of 38° against the stacking layers (Fig. 3), rather than an interdigitation structure. The difference between the d values for the polymer and monomer crystals is because of a change in repeating structure of the stacked diene moieties during polymerization to the 2,5-*trans*-structure accompanied by a rotation of the muconate moiety^{11,12}.

The transformation of the poly(muconic acid) crystals to the ammonium polymer crystals under various conditions is illustrated in Fig. 4. We performed the reactions with a limiting amount of amine to observe the intermediate structure of the crystals in the process of transformation. When the ratio of the amine to the carboxylic acid was low, all the amine molecules added were

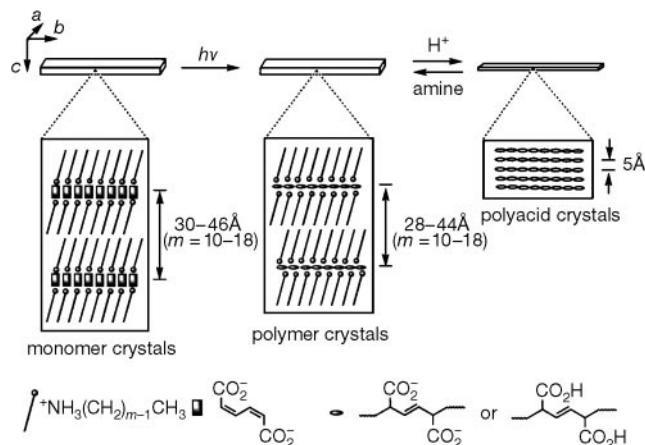


Figure 3 Model of photopolymerization and polymer transformation performed in the crystalline state. The alkylamine molecules and protons are possibly introduced via the *a*–*c* plane of the acid and ammonium polymer crystals (see refs 11,12 for the crystal structures of the related monomer and polymer crystals).

Table 1 Characteristics of the polymer crystals of *n*-alkylammonium (*Z,Z*)-muconates*

<i>m</i> †	Monomer crystals		Polymer crystals‡			Transformed polymer crystals§		
	2θ (°)	d (Å)	Yield (%)	2θ (°)	d (Å)	Conversion (%)	2θ (°)	d (Å)
10	2.99	29.6	66	3.17	27.9	93.6	3.11	28.4
12	2.63	33.6	80	2.79	31.7	94.7	2.74	32.2
14	2.34	37.8	96	2.46	35.5	95.1	2.45	36.1
16	2.11	41.9	100	2.24	39.4	95.3	2.20	40.2
18	1.93	45.8	95	2.02	43.7	95.9	1.99	44.4

* *n*-Alkylammonium (*Z,Z*)-muconates were prepared from (*Z,Z*)-muconic acid and the corresponding amines and recrystallized from methanol or ethanol to yield thin or powdery crystals. The 2θ values were determined by small-angle powder X-ray diffraction (Mac Science D1P1000, camera length 30 cm, graphite-monochromatized Cu K_α line). 2θ values were calibrated with stearic acid.

† The carbon number in the *n*-alkylammonium group.

‡ The photopolymerization was carried out with a high-pressure mercury lamp (Toshiba SHL-100, 100W, Pyrex filter) for 8 h under atmospheric conditions. The polymer crystals were isolated as polymerized after the unreacted monomer was removed by extraction with methanol. The yield was determined gravimetrically.

§ Poly(dodecylammonium muconate) was used as the starting polymer. The hydrolysis of the ammonium polymer was carried out in HCl methanol solution (1 mol l^{-1}) with stirring at room temperature for 30 min. Elemental analysis confirmed the quantitative transformation. The poly(muconic acid) crystals (50 mg) were dispersed in methanol (20 ml), to which the corresponding amine was added, and the solution was stirred at room temperature for 0.5–1 h. The polymer crystals were filtered, washed with methanol, and dried *in vacuo*. The conversion was determined by elemental analysis.

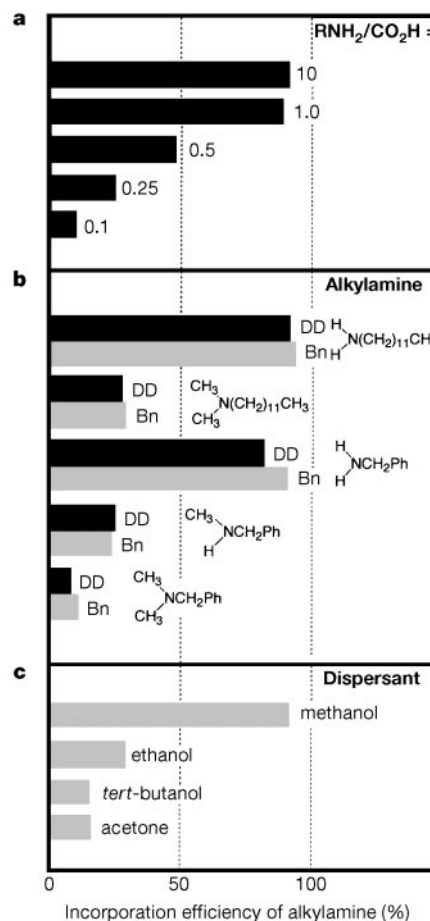


Figure 4 The crystalline-state polymer transformation from poly(muconic acid) to the ammonium polymers. **a**, Effect of the molar ratio of charged dodecylamine to the carboxyl group in the polymer. Dispersant, methanol. **b**, Effect of the structure of alkylamine. Poly(dodecylammonium muconate) (DD) or poly(benzylammonium muconate) (Bn) crystals were used as the precursors of the poly(muconic acid) crystals. Poly(muconic acid) was reacted with dodecylamines or benzylamines. Dispersant, methanol, $\text{RNH}_2/\text{CO}_2\text{H} = 10$. **c**, Effect of dispersant on transformation from poly(muconic acid) to poly(benzylammonium muconate) crystals.

smoothly incorporated into the crystals. Powder X-ray diffraction revealed that the layer structure of alkylammonium is constructed even when only 10% of the amine was introduced, and that the interplanar spacing is equal to the infinite value ($d = 32 \text{ \AA}$). The peak due to poly(muconic acid) crystals ($d = 4.8 \text{ \AA}$) disappeared at 50% conversion (see Supplementary Information). The crystals may contain vacant space in an intermediate state, into which alkylamines are introduced at an accelerating rate in the later stages of the reaction.

The reactions of poly(muconic acid) crystals with both dodecylamine and benzylamine produced a high yield, irrespective of the kind of alkylammonium. When the reactions with secondary or tertiary amine were attempted, they failed, despite an increasing basicity. This indicates an important steric requirement around the polymer chain for the transformation. It also indicates the formation of a hydrogen-bond network between the carboxylate anion and the ammonium cation, which function as the triple hydrogen bond acceptor and donor, respectively¹².

The type of dispersant significantly influenced the conversion, even though the reaction proceeded heterogeneously; as the hydrophobicity of the dispersant increased, the conversion decreased. In the polymer crystals, the polymer chains with a stereoregular structure are aligned along with the *b*-axis of the crystals and linked to each other by the two-dimensional hydrogen-bond network. Both covalent and hydrogen bonds support the reversible crystal-to-crystal transformation. Polar solvent molecules probably help to cut break hydrogen bonds between the carboxylic acids. The cooperative process of the insertion of amine molecules and the formation of restructured layers is essential for the transformation. □

Received 7 December 1999; accepted 14 March 2000.

- Schmidt, G. M. J. Photodimerization in the solid state. *Pure Appl. Chem.* **27**, 647–678 (1971).
- Ramamurthy, V. (ed.) *Photochemistry in Organized and Constrained Media* (VCH, New York, 1991).
- Desiraju, G. R. *Crystal Engineering: The Design of Organic Solids* (Elsevier, Amsterdam, 1989).
- Etter, M. C. Encoding and decoding hydrogen-bond patterns of organic compounds. *Acc. Chem. Res.* **23**, 120–126 (1990).
- Aakeröy, C. B. & Seddon, K. R. The hydrogen bond and crystal engineering. *Chem. Soc. Rev.* **1994**, 283–288.
- Russell, V. A., Evans, C. C., Li, W. & Ward, M. D. Nanoporous molecular sandwiches: pillared two-dimensional hydrogen-bonded networks with adjustable porosity. *Science* **276**, 575–579 (1997).
- Zaworotko, M. J. Cooperative bonding affords a wholesome story. *Nature* **386**, 220–221 (1997).
- Meléndez, R. E. & Hamilton, D. A. Hydrogen-bonded ribbons, tapes and sheets as motifs for crystal engineering. *Top. Curr. Chem.* **198**, 97–129 (1998).
- Rogers, R. D. & Zaworotko, M. J. Whither crystal engineering. *Trans. Am. Crystallogr. Assoc.* **33**, 1–5 (1998).
- Matsumoto, A. *et al.* Crystalline-state polymerization of diethyl (Z,Z)-2,4-hexadienedioate via a radical chain reaction mechanism to yield an ultrahigh-molecular-weight and stereoregular polymer. *Macromolecules* **31**, 2129–2136 (1998).
- Tashiro, K. *et al.* Structure analysis of monomer and polymer crystals in the photoinduced solid-state polymerization reaction of diethyl *cis,cis*-muconate. *Macromolecules* **32**, 7946–7950 (1999).
- Matsumoto, A., Odani, T., Chikada, M., Sada, K. & Miyata, M. Crystal-lattice controlled photopolymerization of di(benzylammonium) (Z,Z)-muconates. *J. Am. Chem. Soc.* **121**, 11122–11129 (1999).
- Matsumoto, A. & Odani, T. Topochemical polymerization of 1-naphthylmethyl ammonium sorbate and characterization of the resulting stereoregular polymer. *Polym. J.* **31**, 717–719 (1999).
- Whittingham, M. S. & Jacobsen, A. J. (eds) *Intercalation Chemistry* (Academic, New York, 1982).
- Coleman, A. W. *et al.* Novel layer structure of sodium calix[4]arenesulfonate complexes: a class of organic clay mimics? *Angew. Chem. Int. Edn Engl.* **27**, 1361–1362 (1988).
- Miyata, M. *et al.* Guest-responsive structural changes in cholic acid intercalation crystals. *Nature* **343**, 446–447 (1990).
- Ogawa, M. & Kuroda, K. Photofunctions of intercalation compounds. *Chem. Rev.* **95**, 399–438 (1995).
- Ogawa, M. & Kuroda, K. Preparation of inorganic-organic nanocomposites through intercalation of organoammonium ions into layered silicates. *Bull. Chem. Soc. Jpn* **70**, 2593–2618 (1997).
- Biradha, K., Dennis, D., MacKinnon, V. A., Shara, C. V. K. & Zaworotko, M. J. Supramolecular synthesis of organic laminates with affinity for aromatic guests: a new class of clay mimics. *J. Am. Chem. Soc.* **120**, 11894–11903 (1998).
- Swift, J. A., Pivovar, A. M., Reynolds, A. M. & Ward, M. D. Template-directed architectural isomers of open molecular frameworks: engineering of crystalline clathrates. *J. Am. Chem. Soc.* **120**, 5887–5894 (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Correspondence and requests for materials should be addressed to A.M. (e-mail: matsumoto@chem.eng.osaka-cu.ac.jp).

Constraining the atmospheric N₂O budget from intramolecular site preference in N₂O isotopomers

Naohiro Yoshida*† & Sakae Toyoda†‡

* Department of Environmental Science and Technology, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama 226-8502, Japan

† CREST Project, Japan Science and Technology Corporation, Kawaguchi, Saitama 332-0012, Japan

Nitrous oxide (N₂O) is an important trace gas in the atmosphere. It is an active greenhouse gas in the troposphere and it also controls ozone concentration in the stratosphere through nitric oxide production¹. One way to trace the geochemical cycle of N₂O is by measuring the natural abundance of stable isotopes, namely ¹⁵N and ¹⁸O (refs 2–15). Here we report the intramolecular distribution of ¹⁵N within the linear NNO molecule, determined by measuring molecular and fragment ions of N₂O on a modified mass spectrometer. This revealed a preference for ¹⁵N at the central N position, or α -site, within N₂O isotopomers (isotope-containing molecules). Moreover, this preference varied significantly throughout the atmosphere. In the troposphere, low α -site preference indicates local emission of N₂O from soils and fossil-fuel combustion, each with distinct isotopomer signatures, which then mixes with background N₂O. In the stratosphere, on the other hand, loss of N₂O is observed as enhanced α -site preference for ¹⁵N, due to fractionation during ultraviolet photolysis of N₂O. We have constructed an atmospheric mass balance of N₂O, incorporating isotopomer abundance, which shows that the intramolecular distribution of ¹⁵N is a parameter that has the potential to increase significantly the resolution with which sources and sinks of N₂O can be identified and quantified in the atmosphere.

There are many uncertainties concerning the global budget of N₂O and the mechanisms involved in its formation and loss in the atmosphere. This is mainly because the concentration and rate of increase of N₂O are relatively low (about 315 parts per billion (p.p.b.) and 0.25% yr⁻¹ at present) and the residence time in the atmosphere is rather long (approximately 120 years), as well as there being a variety of natural and anthropogenic sources¹. Because of the complexity of its cycle, isotopic characterization is expected to have the potential to reduce uncertainties concerning its global cycle.

N₂O is a linear three-atom molecule in which there is one nitrogen atom at the centre and one at the end. There are 12 isotopomers as a result of this asymmetric structure and combination of nitrogen and oxygen isotopes. Among these isotopomers, the following five are sufficiently abundant to measure, because of adequate isotopic abundance: ¹⁴N¹⁴N¹⁶O, ¹⁴N¹⁵N¹⁶O, ¹⁵N¹⁴N¹⁶O, ¹⁴N¹⁴N¹⁷O and ¹⁴N¹⁴N¹⁸O.

Yoshida and Matsuo² have introduced the bulk nitrogen isotope ratio

$$^{15}R^{\text{bulk}} = (^{15}\text{N}^{14}\text{N}^{16}\text{O} + ^{14}\text{N}^{15}\text{N}^{16}\text{O}) / (2^{14}\text{N}^{14}\text{N}^{16}\text{O}) \quad (1)$$

as a key parameter to illustrate the geochemical cycle of N₂O, followed by several observations and simulation experiments^{3–7}. The oxygen isotope ratio

$$^{18}R = ^{14}\text{N}^{14}\text{N}^{18}\text{O} / ^{14}\text{N}^{14}\text{N}^{16}\text{O} \quad (2)$$

was introduced later⁸. Values of ¹⁵R^{bulk} and ¹⁸R have been indepen-

‡ Present address: Department of Environmental Chemistry and Engineering, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama 226-8502, Japan.

dently determined by Kim and Craig⁹, and later simultaneously by several researchers^{10–15}. The $^{15}R^{\text{bulk}}$, however, is determined as the average abundance of two nitrogen isotopomers, and the ^{18}R is determined on the assumption of mass-dependent fractionation of the oxygen isotopes, ^{17}O and ^{18}O . Slight enrichment of ^{17}O relative to ^{18}O in N_2O has been found in the troposphere^{16–18} and in the stratosphere¹⁹, which indicates some unknown mass-independent fractionation processes for oxygen isotopes. The measurements reported in refs 16–19 were of ^{17}O abundance in O_2 prepared from N_2O :

$$^{17}R = ^{14}\text{N}^{14}\text{N}^{17}\text{O}/^{14}\text{N}^{14}\text{N}^{16}\text{O} \quad (3)$$

and this was measured simultaneously with the ^{18}R .

The isotopomers $^{14}\text{N}^{15}\text{N}^{16}\text{O}$ and $^{15}\text{N}^{14}\text{N}^{16}\text{O}$ have an identical mass, so they have been thought to be indistinguishable using conventional mass spectrometric techniques. Here we establish a method for determination of the site preference between two isotopomers by measuring molecular and fragment ions on a mass spectrometer. We denote the centre and end sites of nitrogen as N^α and N^β , respectively. The nitrogen isotope ratio for each site is expressed as the isotopomer ratio as follows:

$$^{15}R^\alpha = ^{14}\text{N}^{15}\text{N}^{16}\text{O}/^{14}\text{N}^{14}\text{N}^{16}\text{O} \quad (4)$$

$$^{15}R^\beta = ^{15}\text{N}^{14}\text{N}^{16}\text{O}/^{14}\text{N}^{14}\text{N}^{16}\text{O} \quad (5)$$

Equation (1) can be rewritten therefore as the average of equations (4) and (5):

$$^{15}R^{\text{bulk}} = (^{15}R^\alpha + ^{15}R^\beta)/2 \quad (6)$$

The fundamental idea and framework of this fragmentation method are briefly described in the Methods section. In addition, we have developed a modification of the continuous-flow type for determining nanomolar levels of N_2O . Stratospheric samples particularly need this drastic reduction in the sample size without

significant decline in the precision. Figure 1 shows a diagram of this arrangement of our experiment. When pure N_2O of a sample size of more than $1\ \mu\text{mol}$ is introduced via a dual inlet, the overall precision of each isotopomer determination is slightly worse in comparison with bulk isotope determination ($\sim 0.02\%$) but better than 0.05% , while that of $\sim 1\ \text{nmol}$ N_2O in the ambient background concentration level through the GC/IRMS (gas chromatograph/isotope-ratio mass spectrometer) system is between 0.4 and 0.7% (better for nitrogen isotopomers and worse for ^{18}O) at present. We expect to be able to improve the precision of isotopomer determination using the continuous-flow-type modification.

Tropospheric samples were collected $\sim 15\ \text{m}$ above ground in the main campus of Nagoya University in 1997, and $\sim 1.5\ \text{m}$ above ground in the Yokohama campus of the Tokyo Institute of Technology in 1998 and 1999. Both sites were urban, close to heavy traffic, and surrounded by patches of secondary forest. Stratospheric samples were collected in balloon-borne cryogenic samplers cooled with liquid He at altitudes between 14.9 and $23.4\ \text{km}$ over Sanriku, Japan (40°N , $142\text{--}143^\circ\text{E}$) on 3 September 1998.

The bulk nitrogen and oxygen isotope ratios of tropospheric N_2O expressed in the δ notation (see Methods) are found to be $7.0 \pm 0.6\%$ for $\delta^{15}\text{N}$ and $43.7 \pm 0.9\%$ for $\delta^{18}\text{O}$ with an average N_2O concentration of $315.7 \pm 2.4\ \text{p.p.b.}$ (concentration was determined from the samples of 1999), which are identical with values recently observed in the troposphere^{10,14}. A large site preference ($\delta^{15}\text{N}^\alpha - \delta^{15}\text{N}^\beta$), however, has been observed in the sense that ^{15}N at the centre (α -site) nitrogen is enriched by $18.7 \pm 2.2\%$ on average compared with the end-site nitrogen (β -site). The tropospheric site preference is larger when $^{15}\text{N}^\beta$ is depleted, as shown in Fig. 2. Background N_2O , for which the site preference is $\sim 19\%$, mixes with locally emitted N_2O both from soil microbiological activities and from fossil-fuel combustion, which has also been implied from $\delta^{17}\text{O}$ and $\delta^{18}\text{O}$ measurements in urban air¹⁸.

By analysing isotopomers, the mixing of background air and end members of two surface sources is seen more clearly than in a conventional bulk ^{15}N – ^{18}O diagram, both in the magnitude ($8\text{--}12\%$ against $2\text{--}3\%$) and in the direction. Of the end members, nitrous oxide from fossil-fuel combustion can be assumed to have small site preference and enriched $^{15}\text{N}^\beta$, whereas N_2O emitted from

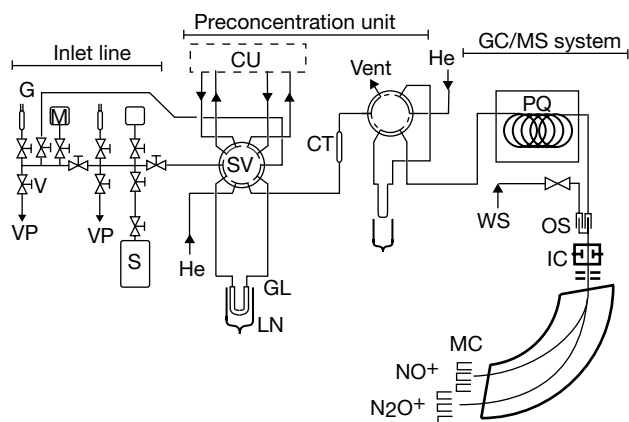


Figure 1 Schematic diagram of the measurement system for isotopomer ratios in N_2O . A $100\text{--}150\ \text{ml}$ aliquot of air in a sample container (S), the N_2O concentration of which ranges from 150 to $350\ \text{p.p.b.}$, is introduced via the inlet line. Nitrous oxide, pre-concentrated and purified through the pre-concentration unit and the gas chromatograph, enters an ionization chamber (IC) of a mass spectrometer with a continuous flow of He, and molecular and fragment ions are monitored on a modified multicollector system (MC). For pure N_2O samples, both sample and working standard are introduced into the mass spectrometer via a conventional dual inlet system (not shown). Other components are: G, vacuum gauge; M, capacitance manometer; V, stainless bellows valve; VP, vacuum pump; GL, stainless tube packed with glass beads; LN, liquid nitrogen baths; CU, unit for concentration measurement; SV, two-position switching valve; CT, chemical trap packed with Ascarite and magnesium perchlorate for removal of CO_2 and H_2O ; PQ, Poraplot Q capillary column ($25\ \text{m}$ long, 25°C isothermal); OS, open-split-type GC/MS interface; WS, working standard (pure N_2O).

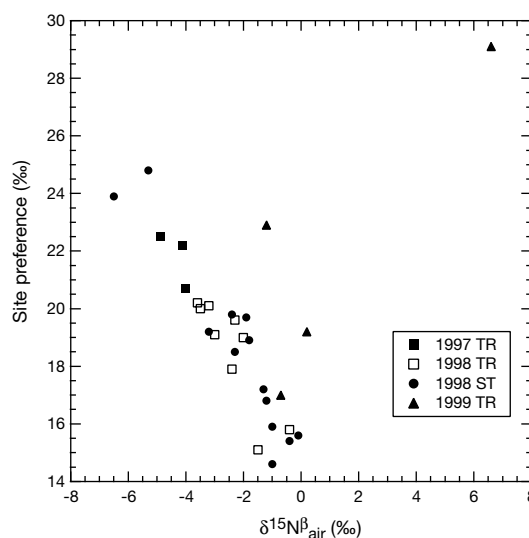


Figure 2 Site preference and nitrogen isotope ratio at the β -site in atmospheric N_2O . 1997 TR, tropospheric samples collected in Nagoya, Japan in 1997. 1998 TR and 1999 TR, tropospheric samples collected in Yokohama, Japan in 1998 and 1999, respectively. 1998 ST, stratospheric samples collected over the northeastern Japan in 1998. Each data point is an average of two or three analyses.

soil would accordingly be enhanced in site preference and depleted in $^{15}\text{N}^\beta$. Thus, isotopomer analysis can potentially distinguish these two surface sources of N_2O , which are less distinguishable on the basis of their bulk ^{15}N and ^{18}O signatures. Furthermore, there would be distinct isotopomer compositions of N_2O derived from nitrification and from denitrification, the two microbial processes that produce N_2O in soils and the oceans. As these processes involve formation and cleavage of the $\text{N}^\alpha\text{--O}$ bond respectively, associated isotopic fractionation will accumulate at the N^α position only, assuming negligible secondary fractionation (such fractionation is observed in atoms adjacent to those involved directly in the reaction, for example, N^β , and is always much smaller than the primary fractionation.) Therefore, $\delta^{15}\text{N}^\alpha$ is expected to vary more than $\delta^{15}\text{N}^\beta$ in the microbiological processes.

The bulk ^{15}N and ^{18}O in stratospheric N_2O collected over Japan in 1998 are enriched at the higher altitude, as shown in Fig. 3a. This enrichment in bulk isotope ratios was expected from observations^{10,14} and from simulation experiments of photolysis at wavelengths of about 200 nm (ref. 20) done in the bulk isotope level. The bulk nitrogen and oxygen isotope ratios are close to those reported by Rahn and Wahlen¹⁴, with the slight difference reflecting

the heterogeneity caused by transportation and physical chemistry in the stratosphere. As is also shown in Fig. 3b, the stratospheric bulk data observed in this study fall on the centre line that is observationally expected¹⁴ using the tropospheric average $\delta^{15}\text{N}^\text{bulk}$ of 7.0‰ that we have found. The remarkable site preference of nitrogen isotopomers in the stratosphere is shown in Fig. 2 and in the top and bottom series of plots in Fig. 3b. This suggests that the site preference found in tropospheric N_2O is enhanced in the stratosphere mainly due to isotopomer fractionation during ultraviolet-induced photolysis that favours the enrichment in ^{15}N at the α -site much more than at the β -site. This is caused by differences in the ultraviolet absorption cross-sections of the isotopomers, as inferred from zero-point energy differences of the ground state among individual isotopomers²¹. The present study offers the first (to our knowledge) confirmation of this effect, via field observation.

The enhancement in the site preference of stratospheric N_2O may partly be due to the reaction with excited oxygen atoms, $\text{O}(^1\text{D})$, which dominates in the lower stratosphere. The stratospheric isotopomer data almost fall on the top and bottom lines in Fig. 3b, which are based on calculated²¹ fractionation factor ratios of $^{14}\text{N}^{15}\text{N}^{16}\text{O}$ and $^{15}\text{N}^{14}\text{N}^{16}\text{O}$ relative to $^{14}\text{N}^{14}\text{N}^{18}\text{O}$, as well as the apparent fractionation of $^{14}\text{N}^{14}\text{N}^{18}\text{O}$ observed in the stratosphere¹⁴. They show, however, slightly smaller site preference in the lower stratosphere, which suggests that the isotopomer fractionation of the reactions with $\text{O}(^1\text{D})$ is smaller than those of apparent photolysis (12.9‰ for ^{18}O and 14.5‰ for $^{15}\text{N}^\text{bulk}$; ref. 14) as experimentally estimated²² for ^{18}O to be 6‰. There would also be a difference in fractionation in the branch reactions of $\text{N}_2\text{O} + \text{O}(^1\text{D}) = 2\text{NO}$ and $\text{N}_2\text{O} + \text{O}(^1\text{D}) = \text{N}_2 + \text{O}_2$. At higher stratospheric altitudes, solar ultraviolet flux increases in the longer-wavelength region, which may favour the larger fractionation factor. This also causes the larger site preference in the higher stratosphere. Several groups, including ours, are conducting ultraviolet photolysis experiments to determine the dependence of fractionation factors of individual isotopomers on ultraviolet wavelength. A detailed scheme could then be constructed by combining transportation and physical chemistry in the stratosphere with the isotopomer data of N_2O .

When we assume a simplified material balance of N_2O in the tropospheric reservoir, neglecting tropospheric sinks, we obtain

$$\Sigma F_{\text{Pi}} = F_{\text{TS}} + \Delta T - F_{\text{ST}} \quad (7)$$

where F represents global N_2O fluxes, specifically of the transportation from the troposphere to the stratosphere (F_{TS}), that from the stratosphere to the troposphere (F_{ST}), those of the terrestrial and oceanic sources including natural and anthropogenic ones (F_{Pi}), and ΔT is the annual tropospheric increase. F_{Pi} includes source terms of natural soil and marine nitrification and denitrification, and those from natural and synthesized fertilizers, cattle breeding, biomass burning, industrial sources including fossil-fuel combustion and nylon production, and so on. If the fractionation factor of the transportation between the troposphere, the stratosphere, and the oceans is assumed to be insignificant, we obtain, taking isotopic balance in the bulk isotope level,

$$^{15}\alpha_{\text{Pi}} \, ^{15}R_{\text{Pi}} F_{\text{Pi}} = ^{15}R_{\text{T}}(F_{\text{TS}} + \Delta T) - ^{15}R_{\text{S}} F_{\text{ST}} \quad (8)$$

$$^{18}\alpha_{\text{Pi}} \, ^{18}R_{\text{Pi}} F_{\text{Pi}} = ^{18}R_{\text{T}}(F_{\text{TS}} + \Delta T) - ^{18}R_{\text{S}} F_{\text{ST}} \quad (9)$$

where α_{Pi} represents the fractionation factor of isotopes in the production processes, and R represents the isotope ratios of source materials (R_{Pi}) and of N_2O in the stratosphere (R_{S}) and in the troposphere (R_{T}). The left superscripts of α and R indicate the bulk nitrogen and oxygen isotopes.

The three-reservoir model of the stratosphere, the troposphere and the oceans² including bulk isotope ratios is revised by including isotopomers of N_2O . The nitrogen isotope mass balance equation

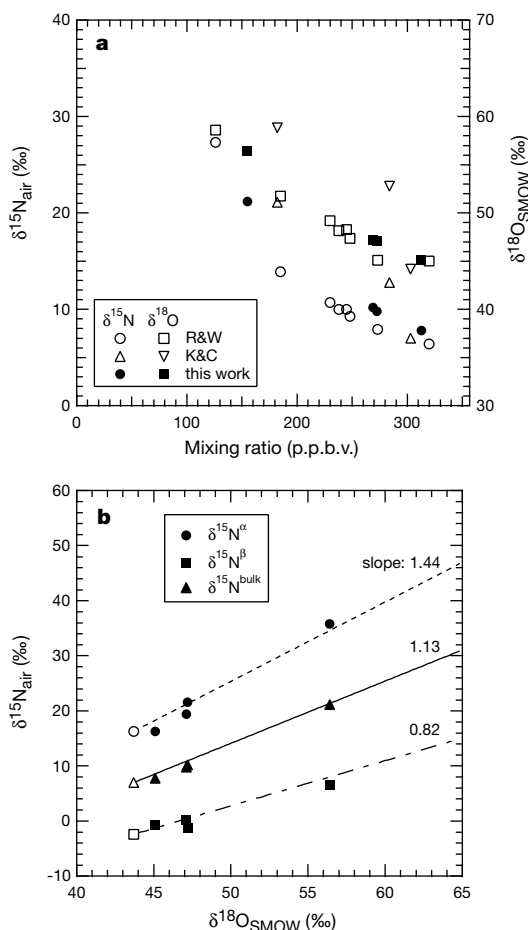


Figure 3 Isotopomer and isotope compositions in stratospheric N_2O . **a**, Bulk nitrogen and oxygen isotope ratios plotted against mixing ratio of N_2O (1 p.p.b.v. = 10^{-9} v/v). Lower mixing ratios correspond to higher altitudes, because N_2O is destroyed by ultraviolet radiation in the stratosphere. Data reported by other researchers (K&C¹⁰, R&W¹⁴) are also plotted. **b**, Relationship between nitrogen isotopomer compositions and oxygen isotope ratio. Most of the data are on the lines, which start from the tropospheric values indicated by open symbols; these lines have slopes that are calculated assuming the apparent ^{18}O enrichment observed in the stratosphere¹⁴, and also assuming the fractionation factor ratio of $^{14}\text{N}^{15}\text{N}^{16}\text{O}$ and $^{15}\text{N}^{14}\text{N}^{16}\text{O}$ to $^{14}\text{N}^{14}\text{N}^{18}\text{O}$ theoretically calculated²¹ for photolysis.

(equation (8)) can be split for each nitrogen isotopomer

$$\Sigma^{15}\alpha_{\text{Pi}}^{15}R_{\text{Pi}}^{\alpha}F_{\text{Pi}} = {}^{15}R_{\text{T}}^{\alpha}(F_{\text{TS}} + \Delta T) - {}^{15}R_{\text{S}}^{\alpha}F_{\text{ST}} \quad (10)$$

$$\Sigma^{15}\alpha_{\text{Pi}}^{\beta}R_{\text{Pi}}^{\beta}F_{\text{Pi}} = {}^{15}R_{\text{T}}^{\beta}(F_{\text{TS}} + \Delta T) - {}^{15}R_{\text{S}}^{\beta}F_{\text{ST}} \quad (11)$$

where the right superscripts of α and R indicate the nitrogen isotopomers. The observation of significant site preference and its large variability in the troposphere and the stratosphere implies that we can more closely constrain the N_2O cycle by isotopomers (equations (10) and (11)) than by bulk isotopes (equation (8)).

The transportation flux of N_2O from the stratosphere to the troposphere across the tropopause, F_{ST} , is estimated to be $\sim 47.2 \text{ Tg N yr}^{-1}$ (ref. 23), ΔT to be 3.9 Tg N yr^{-1} , the stratospheric sink to be $12.3 \text{ Tg N yr}^{-1}$ (ref. 1), and the source term from the terrestrial and marine ecosystems and other anthropogenic sources is estimated to be 16.2 (ranging from 6.4 to 34.4) Tg N yr^{-1} (ref. 24). From these flux values and the stratospheric enriched isotopomer values estimated from the present observation ($\delta^{15}\text{N}^{\alpha} = 22.0\text{‰}$, $\delta^{15}\text{N}^{\beta} = 0.7\text{‰}$ and $\delta^{18}\text{O} = 48.3\text{‰}$), the weighted average isotopomer values of N_2O from the sources are calculated by assuming the mass balance to compose the N_2O isotopomer ratios in the troposphere in equations (9) to (11). Average source terms are calculated to be as follows; $\delta^{15}\text{N}^{\alpha} = -0.3\text{‰}$ (ranging from -25.7 to 8.5‰), $\delta^{15}\text{N}^{\beta} = -11.4\text{‰}$ (ranging from -25.3 to -6.7‰), and $\delta^{18}\text{O} = 30.3\text{‰}$ (ranging from 9.8 to 37.4‰). This gives an estimated site preference of 11.1‰ (ranging between -0.5 and 15.1‰), and an estimated $\delta^{15}\text{N}^{\text{bulk}}$ of -5.8‰ (ranging between -25.5 and 1.0‰). These estimates are for the flux-weighted average of the oceanic and terrestrial sources.

In summary, our present results confirm that the site preference of nitrogen isotopomers in tropospheric N_2O (18.7‰) is the result of the enhancement by back injection from the stratosphere (21.3‰) where isotopomer fractionation during photolysis and oxidation occurs. The 'original' site preference (11.3‰) of tropospheric N_2O should be in accordance with the signature of production and/or consumption processes for N_2O taking place in the terrestrial and the oceanic environments.

We are at present determining the oceanic isotopomer signatures, with which terrestrial and oceanic source strength will be closely examined. Similarly, when natural and anthropogenic sources are isotopomerically characterized, their relative contributions will be assessed in detail. We are also conducting a series of simulation experiments on pure microbial culture, and also field observations of N_2O emitted from different types of soil and environment, so that nitrification and denitrification might be more clearly distinguished. Physical processes such as transportation, mixing and diffusion also need to be carefully studied in field observations, as do the isotopomer effects of other chemical processes²⁵.

Isotopomer analysis of N_2O has produced more fundamental and sensitive information than that available from bulk isotope analysis, and the information has helped the understanding of not only the mechanisms of the producing and consuming pathways, but also their contributions to the global cycle of N_2O . We expect that isotopomer information from process studies and from field observations, together with source and sink strength reassessments, will improve our understanding of the budget of N_2O , the least-understood greenhouse gas. □

Methods

Mass fragmentation analysis

The molecular ion of N_2O is measured first and its fragment ion NO^+ is subsequently measured on a MAT 252 mass spectrometer with Faraday cups specially designed so that the isotope ratios for molecular and fragment ions of N_2O can be determined (Fig. 1). The fragmentation ratio, as expressed in the ratio of the intensity at mass to charge (m/z) values of 30 and 44, is ~ 0.30 with a small fluctuation of $\pm 0.1\%$ (relative to the ratio) when ion source conditions are kept constant. Reproducibility of the isotope ratio determinations of the fragment ion NO^+ is $0.035\text{--}0.087\text{‰}$ (1σ), which is slightly worse than that of the molecular ion N_2O^+ ($0.008\text{--}0.062\text{‰}$) because of the lower ion beam intensity. The

observed ion beam intensity ratios at m/z values of 45, 46 and 31 can be expressed as follows:

$$^{45}R = {}^{15}R^{\alpha} + {}^{15}R^{\beta} + {}^{17}R \quad (12)$$

$$^{46}R = {}^{18}R + ({}^{15}R^{\alpha} + {}^{15}R^{\beta}){}^{17}R + {}^{15}R^{\alpha}{}^{15}R^{\beta} \quad (13)$$

$$^{31}R = {}^{15}R^{\alpha} + {}^{17}R \quad (14)$$

From equations (12) to (14), ${}^{15}R^{\alpha}$, ${}^{15}R^{\beta}$ and ${}^{18}R$ are calculated on the assumption of oxygen mass dependent fractionation:

$$^{18}R/{}^{18}R_{\text{std}} = ({}^{17}R/{}^{17}R_{\text{std}})^2 \quad (15)$$

The ${}^{17}\text{O}$ anomaly being about 1‰ (refs 18, 19) causes a slight overestimate of 0.1 and 0.05‰ for $\delta^{15}\text{N}^{\alpha}$ and $\delta^{15}\text{N}^{\text{bulk}}$, respectively, where the δ s are expressed in ‰ deviation from the standard; $\delta^{15}\text{N}^{\alpha} = 1,000 ({}^{15}R_{\text{sample}}^{\alpha}/{}^{15}R_{\text{std}}^{\alpha} - 1)$, $\delta^{15}\text{N}^{\beta} = 1,000 ({}^{15}R_{\text{sample}}^{\beta}/{}^{15}R_{\text{std}}^{\beta} - 1)$, $\delta^{15}\text{N}^{\text{bulk}} = 1,000 ({}^{15}R_{\text{sample}}^{\text{bulk}}/{}^{15}R_{\text{std}}^{\text{bulk}} - 1)$, $\delta^{17}\text{O} = 1,000 ({}^{17}R_{\text{sample}}/{}^{17}R_{\text{std}} - 1)$, and $\delta^{18}\text{O} = 1,000 ({}^{18}R_{\text{sample}}/{}^{18}R_{\text{std}} - 1)$. The standard for ${}^{15}\text{N}$ is atmospheric N_2 , and that for ${}^{18}\text{O}$ is V-SMOW.

Correction for the stable rearrangement and standardization

Rearrangement or scrambling reactions of N_2O during electron impact ionization have been reported by several researchers^{26,27}. If we define the rearrangement fraction, γ , as the ratio of NO^+ whose nitrogen comes from the β -site to total NO^+ , the observed isotope ratio of NO^+ can be expressed as follows:

$${}^{15}R_{\text{observed}}^{\alpha} = (1 - \gamma){}^{15}R^{\alpha} + \gamma{}^{15}R^{\beta} \quad (16)$$

The relationship between the nitrogen isotope ratios of the original α -site, observed α -site and the bulk is expressed as

$$\delta^{15}\text{N}^{\alpha} = \delta^{15}\text{N}_{\text{observed}}^{\alpha} + 2\gamma A^{-1}(\delta^{15}\text{N}_{\text{observed}}^{\alpha} - \delta^{15}\text{N}^{\text{bulk}})/(1 - 2\gamma) \quad (17)$$

and from equations (6), (16) and (17), we obtain the nitrogen isotope ratio at the β -site

$$\delta^{15}\text{N}^{\beta} = \delta^{15}\text{N}^{\text{bulk}} + A(\delta^{15}\text{N}^{\text{bulk}} - \delta^{15}\text{N}^{\alpha})/(2 - A) \quad (18)$$

where the parameter A is a constant and equals $({}^{15}R^{\alpha}/{}^{15}R^{\text{bulk}})_{\text{std}}$. The rearrangement masks the site preference entirely if γ equals 0.5 . We have found a stable γ value of 0.0811 ± 0.0006 that is in the range of the reported values^{26,27}. This low and stable value of γ means the original site preference can be evaluated from the present fragmentation method. The precision of ${}^{15}\text{N}^{\alpha}$ and ${}^{15}\text{N}^{\beta}$ depend simply on analytical errors of the observed values because the uncertainties of A and γ are negligible and less than 1% . The working standard N_2O gas for the determination of isotopomers has been calibrated with a primary standard that was synthesized by the thermal decomposition of ammonium nitrate whose ${}^{15}\text{N}$ has been determined independently for NH_4^+ and NO_3^- with the conventional method. This reaction has the advantage that the nitrogen atom at the α -site of the produced N_2O originates from NO_3^- and that at the β -site from NH_4^+ (refs 26, 28). The analytical system designed and developed in the present study requires careful settings, the details of which are described elsewhere^{29,30}. The instrumentation and standardization for the determination of nitrogen isotopomers of N_2O have been established. The fragmentation method developed in the present study can be used for the determination of the site preference of isotopes in the isotopomers of other molecular species, such as those including distinguishable sites for carbon atoms. The mass spectrometric determination method is used as a reference for other optical methods that are now in progress.

Received 28 October 1999; accepted 28 March 2000.

- Houghton, J. T. *et al.* (eds) *Climate Change 1995: The Science of Climate Change* (Cambridge University Press, 1995).
- Yoshida, N. & Matsuo, S. Nitrogen isotope ratio of atmospheric N_2O as a key to the global cycle of N_2O . *Geochim. J.* **17**, 231–239 (1983).
- Yoshida, N. *et al.* ${}^{15}\text{N}/{}^{14}\text{N}$ ratio of dissolved N_2O in the eastern tropical Pacific Ocean. *Nature* **307**, 442–444 (1984).
- Yoshida, N. ${}^{15}\text{N}$ -depleted N_2O as a product of nitrification. *Nature* **335**, 528–529 (1988).
- Yoshida, N. *et al.* Nitrification rates and ${}^{15}\text{N}$ abundances of N_2O and NO_3^- in the western North Pacific. *Nature* **342**, 895–897 (1989).
- Yamazaki, T., Yoshida, N., Wada, E. & Matsuo, S. N_2O reduction by *Azotobacter vinelandii* with emphasis on kinetic isotope effects. *Plant Cell Physiol.* **28**, 263–271 (1987).
- Ueda, S., Ogura, N. & Wada, E. Nitrogen isotope ratio of groundwater N_2O . *Geophys. Res. Lett.* **18**, 1449–1452 (1991).
- Wahlen, M. & Yoshinari, T. Oxygen isotope ratios in N_2O from different environments. *Nature* **313**, 780–782 (1985).
- Kim, K. -R. & Craig, H. Two-isotope characterization of N_2O in the Pacific Ocean and constraints on its origin in deep water. *Nature* **347**, 58–61 (1990).
- Kim, K. -R. & Craig, H. Nitrogen-15 and oxygen-18 characteristics of nitrous oxide: A global perspective. *Science* **262**, 1855–1857 (1993).
- Inoue, H. Y. & Mook, W. G. Equilibrium and kinetic nitrogen and oxygen isotope fractionations between dissolved and gaseous N_2O . *Chem. Geol.* **113**, 135–148 (1994).
- Tanaka, N. *et al.* High precision mass spectrometric analysis of isotopic abundance ratios in nitrous oxide by direct injection of N_2O . *Int. J. Mass Spectrom. Ion Proc.* **142**, 163–175 (1995).
- Yoshinari, T. *et al.* Nitrogen and oxygen isotopic composition of N_2O from suboxic waters of the eastern tropical North Pacific and the Arabian Sea – measurement by continuous-flow isotope-ratio monitoring. *Mar. Chem.* **56**, 253–264 (1997).

14. Rahn, T. & Wahlen, M. Stable isotope enrichment in stratospheric nitrous oxide. *Science* **278**, 1776–1778 (1997).
15. Dore, J. E., Popp, B. N., Karl, D. M. & Sansone, F. J. A large source of atmospheric nitrous oxide from subtropical North Pacific surface waters. *Nature* **396**, 63–66 (1998).
16. Thiemens, M. H. & Trogler, W. C. Nylon production: An unknown source of atmospheric nitrous oxide. *Science* **251**, 932–934 (1991).
17. Cliff, S. S. & Thiemens, M. H. High-precision isotopic determination of the $^{18}\text{O}/^{16}\text{O}$ and $^{17}\text{O}/^{16}\text{O}$ ratios in nitrous oxide. *Anal. Chem.* **66**, 2791–2793 (1994).
18. Cliff, S. S. & Thiemens, M. H. The $^{18}\text{O}/^{16}\text{O}$ and $^{17}\text{O}/^{16}\text{O}$ ratios in atmospheric nitrous oxide: A mass-independent anomaly. *Science* **278**, 1774–1776 (1997).
19. Cliff, S. S., Brenninkmeijer, C. A. M. & Thiemens, M. H. First measurement of the $^{18}\text{O}/^{16}\text{O}$ and $^{17}\text{O}/^{16}\text{O}$ ratios in stratospheric nitrous oxide: A mass independent anomaly. *J. Geophys. Res.* **104**, 16171–16175 (1999).
20. Rahn, T., Zhang, H., Wahlen, M. & Blake, G. A. Stable isotope fractionation during ultraviolet photolysis of N_2O . *Geophys. Res. Lett.* **25**, 4489–4492 (1998).
21. Yung, Y. L. & Miller, C. E. Isotopic fractionation of stratospheric nitrous oxide. *Science* **278**, 1778–1780 (1997).
22. Johnston, J. C., Cliff, S. S. & Thiemens, M. H. Measurement of multioxygen isotopic ($\delta^{18}\text{O}$ and $\delta^{17}\text{O}$) fractionation factors in the stratospheric sink reactions of nitrous oxide. *J. Geophys. Res.* **100**, 16801–116804 (1995).
23. Holton, J. R. On the global exchange of mass between the stratosphere and troposphere. *J. Atmos. Sci.* **47**, 392–395 (1990).
24. Kroeze, C., Mosier, A. & Bouwman, L. Closing the global N_2O budget: A retrospective analysis 1500–1994. *Glob. Biogeochem. Cycles* **13**, 1–8 (1999).
25. Zipf, E. C. & Prasad, S. S. Experimental evidence that excited ozone is a source of nitrous oxide. *Geophys. Res. Lett.* **25**, 4333–4336 (1998).
26. Friedman, L. & Bigeleisen, J. Oxygen and nitrogen isotope effects in the decomposition of ammonium nitrate. *J. Chem. Phys.* **18**, 1325–1331 (1950).
27. Begun, G. M. & Landau, L. Mass spectra and metastable transitions in isotopic nitrous oxide. *J. Chem. Phys.* **35**, 547–551 (1961).
28. Richardson, W. S. & Wilson, E. B. Jr The infra-red spectrum of $^{15}\text{N}^{14}\text{NO}$ and the force constants of nitrous oxide. *J. Chem. Phys.* **18**, 694–696 (1950).
29. Toyoda, S. & Yoshida, N. Determination of nitrogen isotopomers of nitrous oxide on a modified isotope-ratio mass spectrometer. *Anal. Chem.* **71**, 4711–4718 (1999).
30. Brenninkmeijer, C. A. M. & Röckmann, T. Mass spectrometry of the intra-molecular nitrogen isotope distribution of environmental nitrous oxide using fragment-ion analysis. *Rapid Commun. Mass Spectrom.* **13**, 2028–2033 (1999).

Acknowledgements

We thank the members of the balloon launch group of The Institute of Space and Astronautical Science, and the cryogenic sampling group, for cooperation in the stratospheric sampling. We thank M. Thiemens for comments that improved an earlier version of the manuscript.

Correspondence and requests for materials should be addressed to N.Y. (e-mail: naoyoshi@depe.titech.ac.jp).

Simulated influences of Lake Agassiz on the climate of central North America 11,000 years ago

S. W. Hostetler*, P. J. Bartlein†, P. U. Clark‡, E. E. Small§ & A. M. Solomon||

* US Geological Survey, Corvallis, Oregon 97333, USA

† Department of Geography, University of Oregon, Eugene, Oregon 97403, USA

‡ Department of Geosciences, Oregon State University, Corvallis, Oregon 97333, USA

§ Department of Earth Sciences, New Mexico Tech, Socorro, New Mexico 87801, USA

|| US Environmental Protection Agency, Corvallis, Oregon 97333, USA

Eleven thousand years ago, large lakes existed in central and eastern North America along the margin of the Laurentide Ice Sheet. The large-scale North American climate at this time has been simulated with atmospheric general circulation models^{1,2}, but these relatively coarse global models do not resolve potentially important features of the mesoscale circulation that arise from interactions among the atmosphere, ice sheet, and proglacial lakes. Here we present simulations of the climate of central and

eastern North America 11,000 years ago with a high-resolution, regional climate model nested within a general circulation model. The simulated climate is in general agreement with that inferred from palaeoecological evidence. Our experiments indicate that through mesoscale atmospheric feedbacks, the annual delivery of moisture to the Laurentide Ice Sheet was diminished at times of a large, cold Lake Agassiz relative to periods of lower lake stands. The resulting changes in the mass balance of the ice sheet may have contributed to fluctuations of the ice margin, thus affecting the routing of fresh water to the North Atlantic Ocean. A retreating ice margin during periods of high lake level may have opened an outlet for discharge of Lake Agassiz into the North Atlantic. A subsequent advance of the ice margin due to greater moisture delivery associated with a low lake level could have dammed the outlet, thereby reducing discharge to the North Atlantic. These variations may have been decisive in causing the Younger Dryas cold event^{3,4}.

We used the GENESIS v2.01 atmosphere general circulation model (AGCM) to simulate the global climate of 11 kyr BP (11,000 years before present). The simulations were run at a nominal atmospheric resolution of 3.75° by 3.75° (latitude by longitude) with 18 vertical layers and an interactive, land–surface physics package^{1,5}. Specified boundary conditions for the AGCM included 11-kyr BP values for eccentricity, obliquity and precession; an atmospheric CO_2 concentration of 270 parts per million by volume, p.p.m.v.; Dorman–Sellers global vegetation⁶; and reconstructions of continental ice sheets 11 kyr ago^{7,8}. Global sea surface temperatures (SSTs) and sea ice were computed by GENESIS using a 50-m mixed-layer ocean model. The GENESIS simulation was initialized from an existing 11-kyr BP simulation and run continuously for 20 years to ensure that the simulated climate and SSTs were in equilibrium. The last five years of the GENESIS simulations were used to derive initial and continuous 12-hour (diurnal) lateral boundary conditions (vertical profiles of temperature, humidity, wind and pressure) for the regional model.

The global climate simulation is similar to others for the period around 11 kyr BP (refs 1, 2) which typically display the interplay of the waning continental ice sheets and rising insolation on Northern Hemisphere atmospheric circulation (Fig. 1). Over the continental interior of North America, the features of the simulated 11-kyr BP climate include: greater summer warming in regions farther from the ice sheet (which occurs in response to the interaction between the large expanse of ice sheet and greater-than-present summer insolation), colder-than-present winter temperatures (which occurs in response to the negative winter insolation anomaly), and precipitation patterns that reflect the reorganization of atmospheric circulation by the ice sheet and insolation. In the AGCM, however, finer-scale topographic detail and surface–atmosphere feedbacks that may be important to climate at regional scales are not resolved.

To resolve these regional-scale feedbacks and processes, we applied the NCAR RegCM2 regional climate model. Our palaeo-climate version of the RegCM2 has an interactive land–surface physics package (the biosphere–atmosphere transfer scheme, BATS) and includes an interactive, one-dimensional (vertical) lake model^{9–13}. A domain size of $3,375 \text{ km} \times 3,825 \text{ km}$, a horizontal grid spacing of $45 \text{ km} \times 45 \text{ km}$ (yielding ~ 80 RegCM2 grid points for each GENESIS grid box) and 17 vertical levels, extending from $\sim 40 \text{ m}$ above the ground to the upper troposphere, were used in the model. A total of 249 lake grid points were prescribed: 185 represent Lake Agassiz (ranging in depth from 10 to 200 m, ref. 14), and the remainder represent other large lakes (all depths set to 25 m). As in the GENESIS simulation, we specified orbital parameters and CO_2 concentration at 11 kyr BP, and a higher-resolution version of the reconstructed Laurentide Ice Sheet (LIS)⁷ was used in the RegCM2.

We constrained the simulated temperatures of the grid points representing Lake Agassiz to be $\leq 5^\circ \text{C}$ based on several lines of evidence. During the time of our simulation, the ostracod *Candona*

subtriangulata, which records sub-surface temperatures and requires cold ($\sim 5^{\circ}\text{C}$) water conditions, overwhelmingly dominated the ostracod assemblage of the Lake Agassiz basin¹⁵. The lake was probably unstratified (polymictic) at 11 kyr BP (ref. 16), so the water column of the open lake probably remained at or below 5°C during ice-free periods. Additional support for cold water temperatures comes from evidence of extensive sediment scraping by lake ice or icebergs¹⁷, estimated meltwater inputs from the LIS of at least 0.02 Sv (ref. 18) and protrusion of the LIS into the lake water along $\sim 1,700$ km of the eastern shoreline which may have formed an ice shelf. Partitioning of the excess heat associated with the constrained temperatures is discussed below. The temperatures of the other large lakes were not constrained.

To quantify the influence of Lake Agassiz on the regional climate, two simulations using RegCM were completed: (1) a continuous 4.5-year-long simulation in which the 11-kyr palaeogeography (that is, LIS, vegetation, proglacial lakes) was specified^{2,19} and (2) a parallel 4.5-year-long simulation in which all lake grid points replaced by the 11-kyr reconstructed vegetation distributions interpolated from the adjacent non-lake grid points. (The first half-year of each simulations is excluded from our analyses to allow for initialization of the model.) We do not include a control simulation for three reasons. First, the model reproduces the present climate of the region adequately^{20,21}. Second, any model bias evident in simulation of the present climate would be small compared with the differences between the 11-kyr simulation and the observed climate data (for example, presence versus absence of the LIS). Third, we compare the simulated climate with that inferred from fossil-pollen data.

The simulated patterns of temperatures and anomalies in both January and July compare well with those inferred for 11 kyr BP from fossil-pollen data (Fig. 2) and reflect the dominant effect of the LIS on the climate of central and eastern North America. Summer temperature gradients as high as 3.3°C per 100 km are located ~ 450 km south of the margin of the LIS, while gradients as high as 5.5°C per 100 km occur closer to the ice sheet, gradients much higher than the present gradient of $\sim 1^{\circ}\text{C}$ per 100 km (ref. 22) and that simulated by the AGCM. A gradient of 3.3°C per 100 km is consistent with the pollen data^{2,23}. The gradient of 5.5°C per 100 km is consistent with reconstructions inferred from fossil-midge data from eastern North America²⁴, but is steeper than that inferred from pollen data, which may reflect the absence of enough pollen sites along the ice sheet to resolve the steeper gradient.

Over central and eastern North America, the magnitude of the moisture anomalies and the location of their gradients are in

generally good agreement with palaeodata^{2,25} (Fig. 2), although in detail, some differences exist between the data and model results. For example, at 11 kyr BP, simulated precipitation is reduced, relative to present, south of Lake Agassiz over central Minnesota; this is consistent with the inference of cooler and drier conditions from pollen data at Elk Lake²⁶, but inconsistent with inferences of wetter conditions inferred from isotopic data at nearby Deep Lake²⁷, attributed to lake-induced precipitation. Additionally, the suggestion of wetter- and warmer-than-present conditions along the southwestern shore of Lake Agassiz may not be comparable to model results because the pollen data from that area present a known 'no-analog' assemblage²⁸. Although our results are in general agreement with palaeodata from central and eastern North America, complete reconciliation of the record of large temporal and spatial transitions in climate that occurred around 11 kyr BP will require detailed compilation of data chronologies and comparison of multiple climate proxies.

Anticyclonic winds at the surface originate from high-pressure centres located over high elevations of the ice sheet (Fig. 3). The flow of the winds is directed downward by the topography of the ice and laterally along the margin by thermal gradients set up between the ice and Lake Agassiz. The simulated anticyclonic winds penetrate a limited distance into adjacent ice-free areas, indicating that anticyclonic flow, as simulated by the RegCM, probably would not have played a significant role in forming regional aeolian features at 11 kyr BP.

From the anomalies between the simulation that included the lakes and the simulation in which the lakes were replaced by vegetation ('lake' minus 'no-lake'), we see that the primary atmospheric effect of Lake Agassiz is the thermal forcing associated with the seasonal cycle of heat storage. This induces changes in atmospheric pressure, circulation (Fig. 3) and moisture (not shown). Circulation changes are evident, for example, in July when, relative to the no-lake simulation, the cold surface of Lake Agassiz modifies atmospheric pressure cells, leading to stronger anticyclonic flow at the surface and a tendency for more cyclonic flow in the mid-troposphere. (Fig. 3). This flow blocks penetration of winds and moisture from the south and west. A related effect of this circulation pattern is a reduction in atmospheric moisture over the lake and the adjacent ice margin; this results from less evaporation from the cold lake surface than evapotranspiration from the warmer vegetation.

The circulation patterns associated with the lake cause a coherent pattern of substantially reduced precipitation that extends eastward along the ice margin, across the Superior Lobe of the ice sheet that dammed eastern outlets of Lake Agassiz and northeastward into the

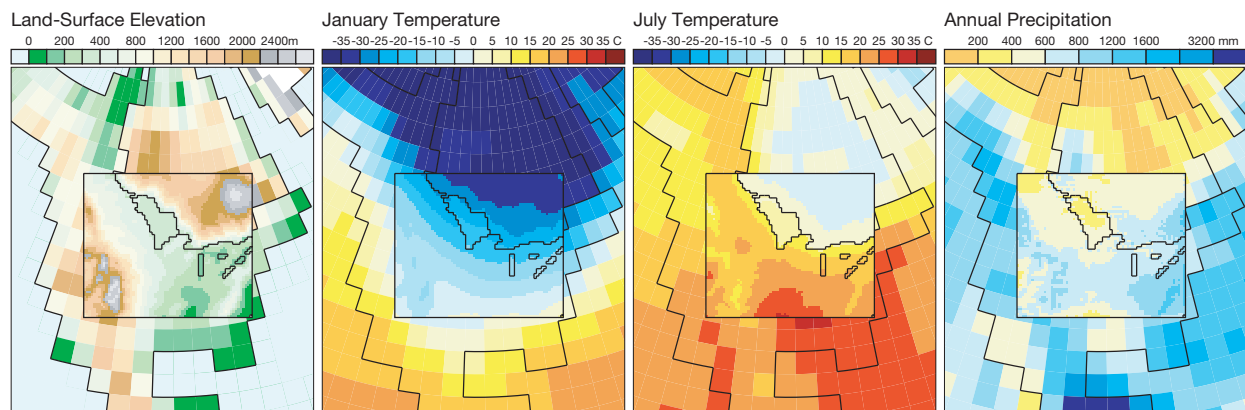


Figure 1 Surface elevation and simulated climate fields over North America at 11 kyr BP. The nominal resolution of the GENESIS atmosphere model, shown as the coarse continental outline and large grid boxes, is 3.75° by 3.75° (400 km by 400 km) latitude by longitude. The resolution of the RegCM2 model used in this study (inset box) is 45 km by 45 km, which yields ~ 80 grid cells for each AGCM grid cell. The AGCM-derived boundary

conditions are assimilated into the RegCM2 every 12 h over an additional 8 grid-point strip, roughly the distance of one AGCM grid cell, along the entire boundary of the model. The figure illustrates that the patterns simulated by the RegCM2 are not simple interpolations of the patterns in the AGCM, but instead reflect the variations of topography and surface characteristics on the higher-resolution grid of the model.

interior of the ice sheet (Fig. 3). Annual precipitation (Fig. 2) is reduced by >120 mm ($\sim 30\%$) over portions of the ice sheet and the Superior Lobe, whereas July precipitation (not shown) is reduced by >50 mm ($\sim 90\%$) around the lake and by 40 mm ($\sim 50\%$) over the ice margin and into the interior. [We note that precipitation is enhanced somewhat by atmospheric heating and elevated moisture levels before the lake freezes in late autumn (November in Fig. 3) but this is offset by drying during the rest of the year.] This generally negative precipitation feedback opposes previous hypotheses^{27,29}, and contrasts with simulations of the US Great Basin in which lake-atmosphere feedbacks enhanced precipitation during the Last Glacial Maximum¹³.

Heat in excess of that required to maintain the lake at a temperature of 5°C (derived mainly from solar heating at the lake surface) would probably have been consumed in warming cold melt

water from the LIS and in melting icebergs and shelf ice. The heat required to raise the estimated inflow of 0.02 Sv of melt water (ref. 18) from 0°C to 5°C is $\sim 10^{13}$ MJ. Accounting for heat needed to warm the melt water, at an ice temperature of -5°C enough excess heat is available from the constrained temperatures ($\sim 10^{14}$ MJ) to melt $\sim 2,500\text{ km}^3$ of ice annually, suggesting that the internal heating of the lake would have contributed to significant ablation of the ice in contact with the lake.

Our results suggest a mechanism for linking the atmosphere-lake-ice system to the retreat and advance of the Superior Lobe, which, in turn, played an important role during the last deglaciation in damming eastward-draining outlets of Lake Agassiz, and thus in controlling the routing of a substantial fraction of continental runoff to the North Atlantic^{3,4,18}. During the large (Lockhart) phase of Lake Agassiz, (which pre-dates eastward drainage to the St Lawrence before 13 kyr BP), the lake would have suppressed precipitation and contributed to retreat of the ice margin in contact with lake water, reinforcing deglaciation. At ~ 13 kyr BP, the Superior Lobe had receded far enough to allow the lake to drain eastward to the St Lawrence River, supplying $\sim 0.05\text{ Sv}$ of additional fresh water to the North Atlantic¹⁸ which triggered the Younger Dryas cold interval^{3,4}. With the lake diminished in size, moisture delivery to the ice sheet would have increased, supporting the hypothesis that the Marquette ice advance into the Superior basin was related to changes in mass balance rather than to surging or ice streaming³⁰. The resulting damming of the eastern outlets would have ended the Younger Dryas cold reversal^{3,4} and allowed the lake to expand to its maximum size (Emerson phase) simulated here.

Our results depend on using the geological record as guidance in limiting the maximum surface temperature of Lake Agassiz, and on the driving boundary conditions derived from GENESIS. In climate simulations, as in nature, the climate of a 5-year period may vary from that of other 5-year periods. The climate forcing arising with the interplay of the LIS and insolation is robust in controlling both the simulated large-scale (AGCM) and regional-scale (RegCM2) circulations. Taken together with the strong thermal forcing of the lake in RegCM2, our results would probably be reproducible for any different period of our GENESIS simulation, even if the actual temperature of the lake was several degrees warmer than suggested

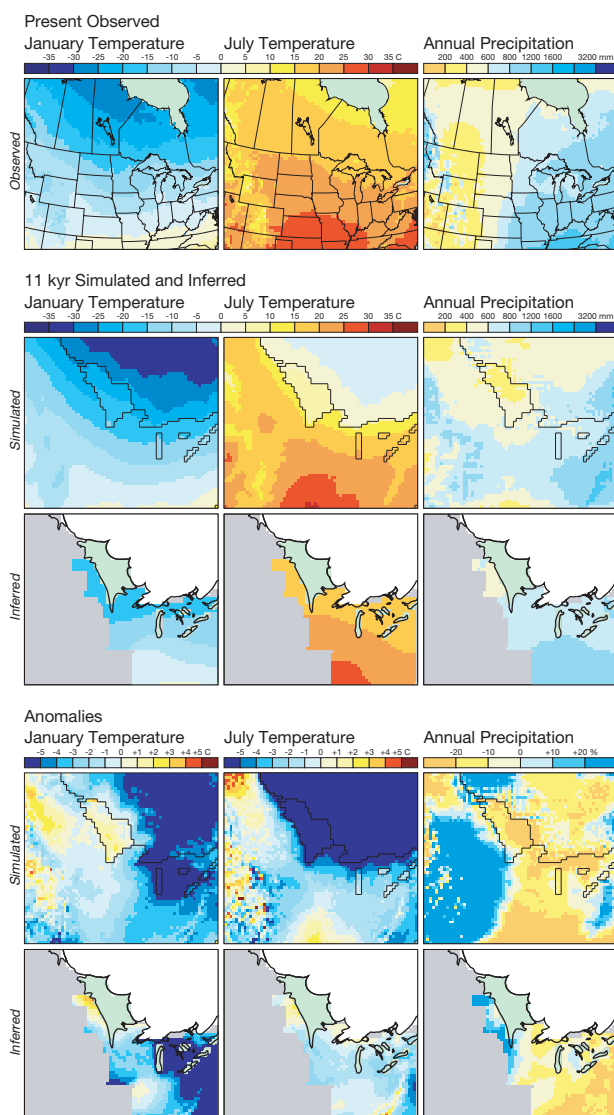


Figure 2 Simulated and inferred temperature and precipitation climatologies for 11 kyr BP. These were inferred from pollen reconstructions and simulated by the RegCM2. The anomalies are the differences between inferred or simulated data relative to observed data. Observed values are 30-year averages from instrumental records interpolated onto the model grid. Simulated values are averages over four simulated years. Inferred values are pollen-based estimates derived by averaging reconstructions for $12,000^{14}\text{C yr BP}$ (~ 14 kyr) and $9,000^{14}\text{C yr BP}$ (~ 10 kyr) (ref. 24) and interpolating the averages onto the model grid.

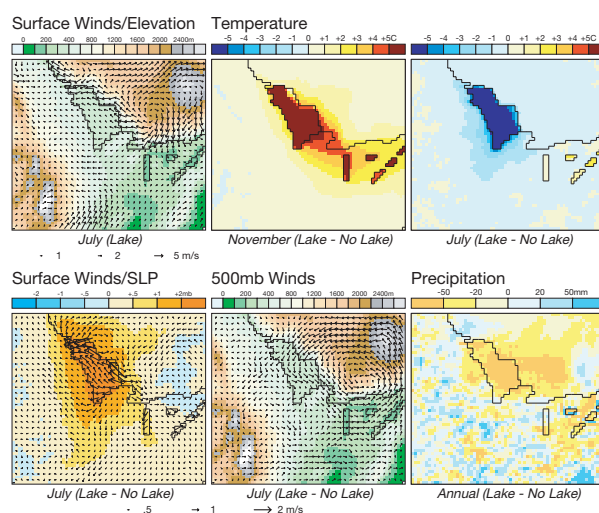


Figure 3 Selected fields for the lake simulation, and anomalies between the lake and no-lake simulations. The values are averages over four simulated years. July surface (2-m) wind vectors and upper-level (500 mb) wind anomalies are plotted over the model topography. Temperature anomalies are for the 2-m air temperature. November is the month in which maximum warming of the atmosphere by the lake occurs. July surface-wind anomalies are plotted over sea-level pressure anomalies (SLP).

by the ostracod data. The ability to simulate finer-scale climate processes in complex palaeoenvironmental settings offers the opportunity to test climate hypotheses at the landscape scale; it also suggests that a nested modelling approach will be useful in investigating a wide range of surface-atmosphere feedbacks associated with regional climates of the past, present and future. □

Received 23 July 1999; accepted 27 March 2000.

- Pollard, D. *et al.* Climate simulations for 10000 and 6000 years BP using the GENESIS global climate model. *Paleoclimates* **2**, 183–218 (1998).
- Bartlein, P. J. *et al.* Paleoclimate simulations for North America over the past 21,000 years: Features of the simulated climate and comparisons with paleoenvironmental data. *Quat. Sci. Rev.* **17**, 549–585 (1998).
- Broecker, W. S. *et al.* Routing of meltwater from the Laurentide ice sheet during the Younger Dryas cold episode. *Nature* **341**, 318–321 (1989).
- Teller, J. T. Meltwater and precipitation runoff to the North Atlantic, Arctic, and Gulf of Mexico from the Laurentide ice sheet and adjacent regions during the Younger Dryas. *Paleoceanography* **5**, 897–905 (1990).
- Thompson, S. L. & Pollard, D. A global climate model (GENESIS) with a land-surface-transfer scheme (LSX), 1, Present-day climate. *J. Clim.* **8**, 732–761 (1995).
- Dorman, J. L. & Sellers, P. J. A global climatology of albedo, roughness length, and stomatal resistance for atmospheric general circulation models as represented by the simple biosphere model (SiB). *J. Appl. Meteorol.* **28**, 833–855 (1989).
- Licciardi, J. M., Clark, P. U., Jenson, J. W. & MacAyeal, D. R. Deglaciation of a soft-bedded Laurentide ice sheet. *Quat. Sci. Rev.* **17**, 427–448 (1998).
- Peltier, W. R. Ice age paleotopography. *Science* **265**, 195–201 (1994).
- Giorgi, F., Marinucci, M. R. & Bates, G. T. Development of a second-generation regional climate model (RegCM2), I, Boundary layer and radiative processes. *Mon. Weath. Rev.* **121**, 2794–2813 (1993).
- Giorgi, F., Marinucci, M. R. & Bates, G. T. Development of a second-generation regional climate model (RegCM2), II, Convective processes and assimilation of boundary conditions. *Mon. Weath. Rev.* **121**, 2814–2832 (1993).
- Hostetler, S. W. & Bartlein, P. J. in *Mechanism of Global Climate Change at Millennial Time Scales* (eds Clark, P. U., Webb, R. S. & Keigwin, L.) 313–328 (AGU Monograph Series, American Geophysical Union, Washington, DC, 1999).
- Small, E. E., Sloan, L. C., Hostetler, S. W. & Giorgi, F. Simulating the water balance of the Aral Sea with a coupled regional climate-lake model. *J. Geophys. Res.* **104**, 6583–6602 (1999).
- Hostetler, S. W., Giorgi, F., Bates, G. T. & Bartlein, P. J. Lake-atmosphere feedbacks associated with paleolakes Bonneville and Lahontan. *Science* **263**, 665–668 (1994).
- Mann, J. D., Leverington, J. D., Rayburn, J. & Teller, J. T. The volume and paleobathymetry of glacial Lake Agassiz. *J. Paleolimnol.* **22**, 71–89 (1999).
- Curry, B. B. Paleochemistry of Lakes Agassiz and Manitoba based on ostracodes. *Can. J. Earth Sci.* **34**, 699–708 (1997).
- Mann, J. D. *et al.* Calculating the volume and heat budget of glacial Lake Agassiz. *GSA Ann. Meeting Abstr.* **29**, A111 (1997).
- Clayton, L., Moran, S. R. & Bluemle, J. P. Intersecting minor lineations on Lake Agassiz plain. *J. Geol.* **73**, 652–656 (1965).
- Licciardi, J. M., Teller, J. T. & Clark, P. U. in *Mechanism of Global Climate Change at Millennial Time Scales* (eds Clark, P. U., Webb, R. S. & Keigwin, L.) 177–202 (AGU Monograph Series, American Geophysical Union, Washington, DC, 1999).
- Dyke, S. & Prest, V. K. *Map 1703A* (Geological Survey of Canada, Ottawa, 1987).
- Bates, G. T., Hostetler, S. W. & Giorgi, F. Two-year simulation of the Great Lakes region with a coupled modeling system. *Mon. Weath. Rev.* **123**, 1505–1522 (1995).
- Giorgi, F. *et al.* Regional nested model simulations of present day and 2XCO₂ climate over the Great Plains of the US. *Clim. Change* **40**, 457–493 (1998).
- Hostetler, S. W. Climate near to the edge of an ice sheet. *Nature* **385**, 393–394 (1997).
- Webb, T., Bartlein, P. J., Harrison, S. P. & Anderson, K. H. in *Global Climates Since the Last Glacial Maximum* (eds Wright, H. E., Kutzbach, J. E., Webb, T., Ruddiman, W. F., Street-Perrott, F. A. & Bartlein, P. J.) 415–467 (Univ. Minnesota Press, 1993).
- Levesque, A. J., Cwynar, L. C. & Walker, I. R. Exceptionally steep north-south gradients in lake temperature during the last deglaciation. *Nature* **385**, 423–426 (1997).
- Laird, K. R., Fritz, S. C., Grimm, E. C. & Mueller, P. G. Century-scale paleoclimatic reconstruction from Moon Lake, a closed-basin lake in the northern Great Plains. *Limnol. Oceanogr.* **41**, 890–902 (1996).
- Bartlein, P. J. & Whitlock, C. in *Elk Lake, Minnesota: Evidence for Rapid Climate Change in the North-Central United States* (eds Bradbury, J. P. & Dean, W. E.) 275–293 (Geologic Society of America, Boulder, 1993).
- Hu, F. S., Wright, H. E., Ito, E. & Lease, K. Climate effects of glacial Lake Agassiz in the midwestern United States during the last deglaciation. *Geology* **25**, 207–210 (1997).
- Overpeck, J. T., Webb, R. S. & Webb, T. W. Mapping eastern North American vegetation change of the past 18 ka: no-analogs and the future. *Geology* **20**, 1071–1074 (1992).
- Teller, J. T. in *North America and Adjacent Oceans During the Last Deglaciation* (eds Ruddiman, W. F. & Wright, H. E.) 39–69 (Geological Society of America, Boulder, 1986).
- Lowell, T. V., Larson, G. H., Hughes, J. D. & Denton, G. H. Age verification of the Lake Gribben forest bed and the Younger Dryas advance of the Laurentide Ice Sheet. *Can. J. Earth Sci.* **36**, 383–393 (1999).

Acknowledgements

We thank F. S. Hu, E. Ito, and H. E. Wright for discussions, D. Pollard and J. Teller for discussions and critical reviews, and R. Gallimore, F. Giorgi and D. Zahnle for assistance. Model simulations were conducted at NCAR and the National Environmental Super Computer (NESC) facility of the US Environmental Protection Agency. Special thanks to NESC for help in completing our model simulations. This work was supported by the US

Geological Survey (S.W.H.), the National Science Foundation (P.J.B., P.U.C., E.E.S.), and the US Environmental Protection Agency (A.M.S.).

Correspondence and requests for materials should be addressed to S.W.H. (e-mail: steve@ucar.edu).

Mesozoic plate-motion history below the northeast Pacific Ocean from seismic images of the subducted Farallon slab

Hans-Peter Bunge* & Stephen P. Grand†

* Department of Geosciences, Princeton University, Princeton, New Jersey 08544, USA

† Department of Geological Sciences, University of Texas, Austin, Texas 78712, USA

The high-resolution seismic imaging of subducted oceanic slabs^{1,2} has become a powerful tool for reconstructing palaeogeography³. The images can now be interpreted quantitatively by comparison with models of the general circulation of the Earth's mantle⁴. Here we use a three-dimensional spherical computer model of mantle convection^{5,6} to show that seismic images of the subducted Farallon plate provide strong evidence for a Mesozoic period of low-angle subduction under North America. Such a period of low-angle subduction has been invoked independently to explain Rocky Mountain uplift far inland from the plate boundary during the Laramide orogeny⁷. The computer simulations also allow us to locate the largely unknown Kula–Farallon spreading plate boundary, the location of which is important for inferring the trajectories of 'suspect' terrain across the Pacific basin⁸.

The geology of western North America was shaped by interaction with three major ocean plates; the Kula, Farallon and Pacific plates⁹. Their motion in the Cenozoic era has been inferred through careful reconstructions of relative plate movement^{10,11}. However, reconstructing Kula and Farallon plate motion further back in time is difficult because the Pacific plate from the Mesozoic era has been almost entirely lost to subduction. The lack of old oceanic lithosphere has prompted attempts to use high-resolution tomographic models to constrain past plate motion³. For example, fast seismic velocity anomalies in the upper mantle shear-wave structure under North America have been related explicitly to Cenozoic subduction of the Farallon plate about 30 Myr ago¹².

Earlier Mesozoic motion (about 80 Myr ago) of the Kula and Farallon plates is now recorded in the lower mantle. In Fig. 1a we show the seismic velocity structure under North America at four depths. The images are from a recent high-resolution tomographic S-wave model¹ and are remarkably similar to an independent study based on P waves². A band of high seismic velocities underlies eastern North America and marks the ancient Kula and Farallon plates.

The 1,300 km snapshot reveals two distinct trends for the slab beneath the Great Lakes region. It strikes southeast under the Canadian shield, but turns sharply south beneath the east coast of the United States. The seismic bight is strongest at 1,500-km depth, where it takes the form of a big 'hook' similar to a feature imaged recently in the deep mantle beneath the Siberian craton³. Under South America the tomographic model reveals a slab at 900-km and 1,100-km depth, but not below.

Geodynamic heterogeneity structure under North America

inferred from mantle circulation is shown in Fig. 1b. It was obtained by imposing the history of Mesozoic–Cenozoic plate motion¹³ as a velocity boundary condition on a three-dimensional (3D) spherical computer simulation of mantle convection^{5,6}. In calculating flow we assumed the mantle viscosity increases by a factor of 100 between the upper and the lower mantle, as indicated by studies of post-glacial rebound¹⁴, the geoid¹⁵ and Earth's rotation¹⁶. We also assumed the mantle is heated primarily from within by radioactivity in agreement with geochemical constraints on the abundance of radioactive elements in the mantle¹⁷, and we included a modest 12% component of heating from the core¹⁸. These assumptions roughly constitute a 'standard model' for whole-mantle convection¹⁹, and their physical effects are quite well understood from previous convection modelling^{6,20,21}.

For comparison with the S-wave study, we applied a tomographic filter²² to the circulation model in Fig. 1b. The filter was calculated by mapping temperature variations into a 3D seismic slowness (inverse velocity) model. For simplicity we assumed a constant temperature to slowness conversion factor of $2 \times 10^{-6} \text{ s km}^{-1} \text{ K}^{-1}$, in line with the available experimental data on silicates in the lower mantle²³. Travel-time anomalies computed from the slowness model were inverted using the algorithm and ray geometry of the S-wave study.

Cold downwelling material from past subduction dominates the geodynamic model at all four depths (Fig. 1b). Under North America the slab runs from the southern Gulf of Mexico to Alaska, following a line that marks the boundary of North America with the Kula and Farallon plates between 80 Myr and 50 Myr ago¹¹.

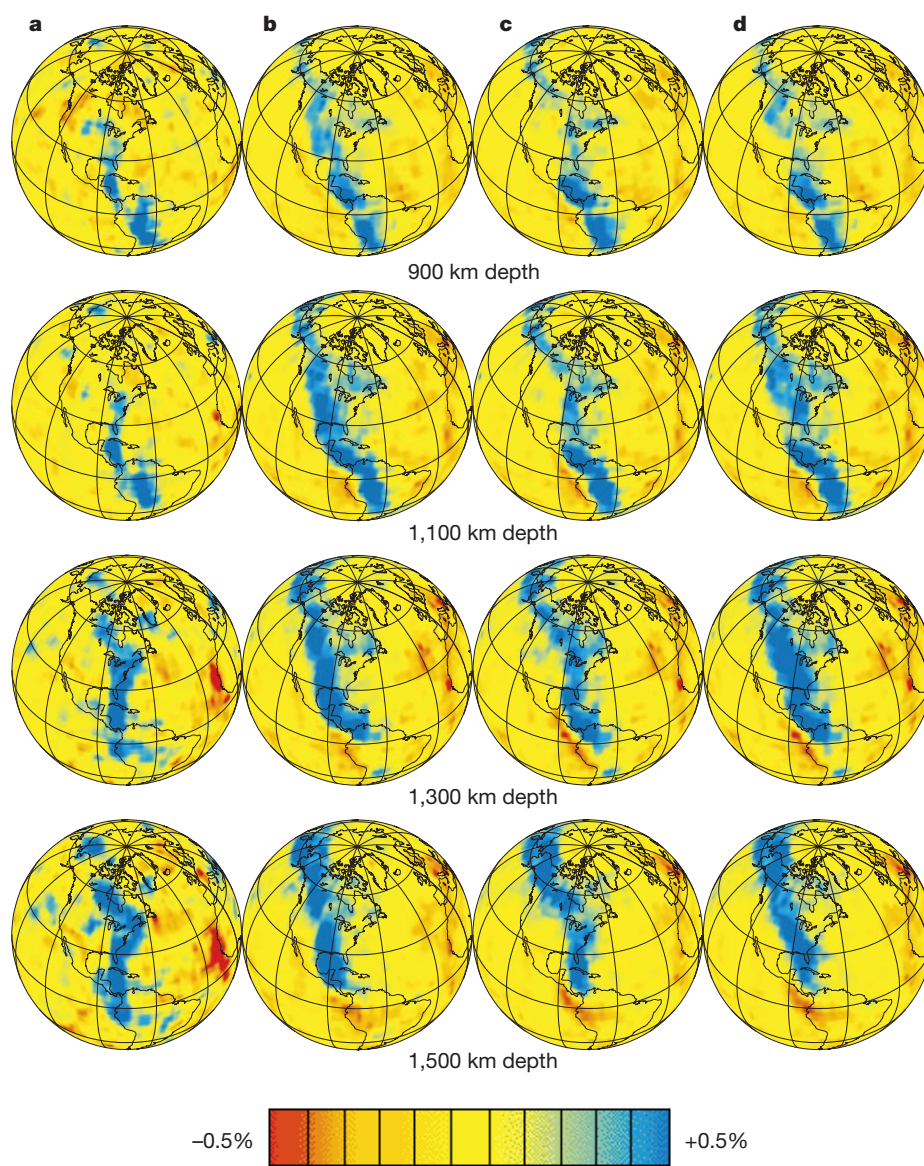


Figure 1 Heterogeneity structure (with mean removed) at 900 km, 1,100 km, 1,300 km and 1,500 km depth. **a**, The seismic shear-wave model of Grand *et al.*¹ **b**, The reference geodynamic model with normal high-angle subduction. **c**, **d**, Two alternative geodynamic models assuming low-angle Laramide subduction and a northern (**c**) or southern (**d**) location for the unknown Kula–Farallon spreading ridge are also shown. All geodynamic models are filtered seismically by inverting their synthetic travel-time anomalies based on the ray-paths and inversion method of the S-wave study (see text). Blues indicate faster than average, and reds slower than average, seismic velocity. In the seismic model a slab-

like feature under the east coast of the United States turns sharply west under the Great Lakes region. In the reference geodynamic model a subducted slab is located under the Rocky Mountain region and runs from the Gulf of Mexico to Alaska. The difference between the geodynamic and seismic model is probably because of shallow-angle subduction of the Farallon plate (see text). The geodynamic models with low-angle Laramide subduction show a slab closer to its location in the seismic model, and the model assuming a northern location of the Kula–Farallon plate boundary correctly predicts the bending of the slab near the Great Lakes region.

Under South America, where the plate reconstructions indicate that subduction began about 60 Myr ago, the simulation reveals a slab at 900 km and 1,100 km but not below.

The thermal anomaly under North America decreases as depth decreases from 1,500 km to 900 km in the lower mantle. A smaller thermal anomaly at the shallower levels is expected, because relatively young ocean floor subducted under North America when the palaeo East Pacific Rise began approaching the continent about 50 Myr ago¹⁰.

Our results indicate the general character of mantle heterogeneity is reproduced well by circulation modelling. However, there is difficulty in reconciling the location of the Farallon slab with its seismic image about 1,500 km further east. Does this difference result from inaccuracies in the plate-motion input model? One of the largest uncertainties in reconstructing North American subduction history is caused by the unusual geological events surrounding the Laramide orogeny. The Cretaceous–Eocene Laramide orogeny (between 70 and 40 Myr ago) caused widespread faulting and uplift of crystalline basement throughout the Rocky Mountain region as far inland as 1,500 km from the western margin of North America²⁴. Successful geological models have invoked low-angle subduction beneath the continent to explain the remarkable landward migration of tectonic activity^{7,25}. Today, shallow-angle subduction occurs in segments of the Nazca plate under central Chile, and may explain the Quaternary period uplift of the Sierras Pampeanas ~700 km inland from the South American/Nazca plate boundary in western Argentina^{26,27}.

Figure 1c shows the tomographic image of a geodynamic model identical to the simulation in Fig. 1b except that we assumed low-angle subduction of the Farallon plate during the Laramide orogeny. In making this assumption we extended the eastern boundary of the ocean plate to account for its landward migration under North America. We used maps that account for the uplift and timing of volcanism in the western United States²⁸ to constrain the regions affected by the flat-lying slab. We kept the relatively northern location of the Kula–Farallon spreading centre implicit in the plate-motion input model¹³ and ignored shallow-angle subduction for the Kula plate, in accordance with its predominantly northward motion under continental North America¹¹. The eastern extent of

shallow-angle subduction is illustrated in Fig. 2.

Changing the plate-motion input model has a profound effect. First, Laramide subduction results in a remarkable agreement between the geodynamic and the tomographic structure, accounting for the slab location under eastern North America. Second, and perhaps more surprisingly, Laramide subduction produces a slab-bend beneath the Great Lakes region where it separates the Kula from the Farallon slab in our model simulation. The proximity of the bend to the seismic bight in Fig. 1a provides a posteriori support for the northern Kula–Farallon plate-boundary option adopted in the plate-motion input model¹³.

The apparent sensitivity of geodynamic heterogeneity structure to the position of the Kula–Farallon plate boundary is of interest because the location of this ancient spreading centre is not well known. Figure 3 shows two vastly different reconstructions that have been proposed, based on the limited information available onshore¹¹. The northern option places the Farallon plate adjacent to the western United States from 85 to 56 Myr ago and produces rapid plate convergence along the continental margin. The southern option moves the Kula–Farallon plate boundary about 3,000 km further south, and the Kula plate lies adjacent to the west coast of the United States. In this option the rapid northern motion of the Kula plate provides an attractive mechanism to carry ‘suspect’ terranes northward along the continental margin⁸.

To address the question of the unknown Kula–Farallon plate boundary, we computed a second Laramide subduction model where we adopted the southern option of the Kula–Farallon spreading ridge. The tomographic image of this model is shown in Fig. 1d. As expected, moving the boundary between the Kula and the Farallon plate about 3,000 km south also moves the slab-bend southward to beneath Florida. South of the bend there is also some eastward migration of the slab compared to the reference simulation. However, further to the north, eastward migration is less pronounced because of the rapid northward motion of the Kula plate.

This example illustrates the potential utility of mantle-circulation

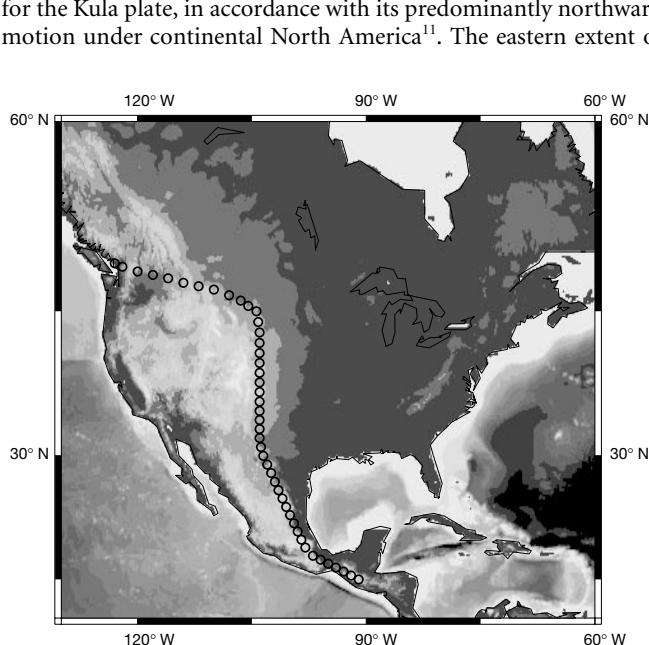


Figure 2 Topographic map of North America where the estimated eastern extent of regions affected by low-angle Farallon subduction is marked with a line of circles²⁸. Throughout Mexico and the western United States the boundary of low-angle subduction coincides with regions of topographic high-stand. Near the Canadian border the line turns sharply west. Assuming the northern location of the Kula–Farallon spreading ridge the turn marks the transition from Kula to Farallon subduction.

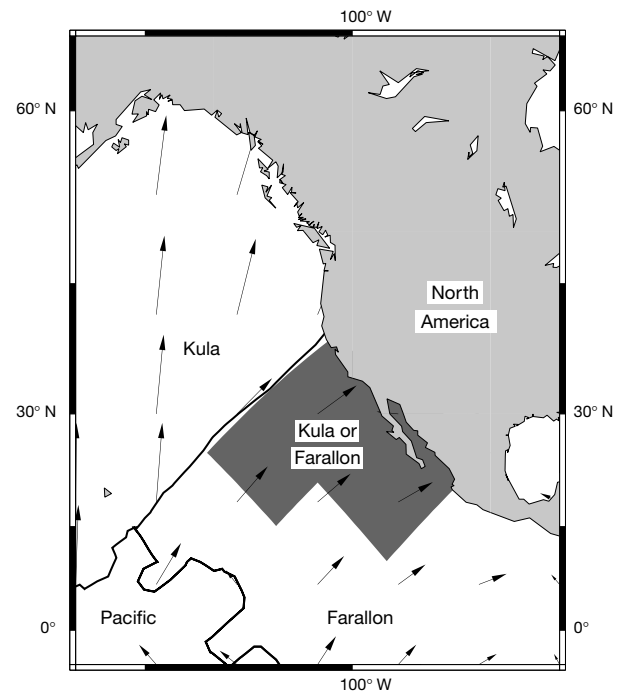


Figure 3 Palaeogeographic map of the northeast Pacific for the time between 64 Myr and 74 Myr ago¹¹. The shaded area indicates the range of possible locations for the Kula–Farallon spreading centre. Note the uncertainty of more than 3,000 km in locating the Kula–Farallon plate boundary.

models in testing plate-tectonic hypotheses with seismic data. Although we cannot test the mechanism for slab flattening^{29,30}, the computer simulations indicate that an uncertainty of 1,500 km in locating ancient plate-boundaries can be detected in the current generation of geodynamic and seismic mantle models. Comparable uncertainty exists in locating other ancient plate-boundaries. For example, the position of the ancient Izanagi subduction zone in the northwest Pacific is largely unknown¹³, and results in small but significant differences in geodynamic and seismic heterogeneity structure under northeastern Asia⁴.

Our simulations still have a long way to go towards realism, including the need for treatment of subduction zones as one-sided downwellings rather than simple fluid drips at convergence zones³¹. Other than lacking plates, the main shortcoming of our models is the absence of horizontal viscosity variations resulting from thermal (or stress) variations. Apart from the piecewise rigidity of the upper boundary layer itself, it is not clear how strongly lateral viscosity variations in the deep mantle affect convection³², and whether our main conclusions would be changed by modelling them. This question needs to be addressed in future studies. □

Received 20 December 1999; accepted 15 March 2000.

- Grand, S. P., van der Hilst, R. D. & Widiyantoro, S. Global seismic tomography: a snapshot of convection in the Earth. *GSA Today* **7**, 1–7 (1997).
- van der Hilst, R. D., Widiyantoro, S. & Engdahl, E. R. Evidence for deep mantle circulation from global tomography. *Nature* **386**, 578–584 (1997).
- van der Voo, R., Spakman, W. & Bijwaard, H. Mesozoic subducted slabs under Siberia. *Nature* **397**, 246–249 (1999).
- Bunge, H.-P. *et al.* Time scales and heterogeneous structure in geodynamic earth models. *Science* **280**, 91–95 (1998).
- Bunge, H.-P. & Baumgardner, J. R. Mantle convection modeling on parallel virtual machines. *Comput. Phys.* **9**, 207–215 (1995).
- Bunge, H.-P., Richards, M. A. & Baumgardner, J. R. A sensitivity study of three-dimensional spherical mantle convection at 10(8) Rayleigh number: Effects of depth-dependent viscosity, heating mode, and an endothermic phase change. *J. Geophys. Res.* **102**, 11991–12007 (1997).
- Coney, P. J. & Reynolds, S. T. Cordilleran Benioff zones. *Nature* **270**, 403–406 (1977).
- DeBiche, M. G., Cox, A. & Engbreton, D. The motion of allochthonous terranes across the North Pacific basin. *GSA Spec. Pap.* **207**, 1–49 (1987).
- Atwater, T. Implications of plate tectonics for the Cenozoic tectonic evolution of western North America. *Geol. Soc. Am. Bull.* **81**, 3513–3536 (1970).
- Stock, J. & Molnar, P. Uncertainties and implications of the Late Cretaceous and Tertiary position of North America relative to the Farallon, Kula and Pacific plates. *Tectonics* **7**, 1339–1384 (1988).
- Engbreton, D. C., Cox, A. & Gordon, R. G. Relative motions between oceanic and continental plates in the Pacific basin. *GSA Spec. Pap.* **206**, 1–59 (1986).
- Van der Lee, S. & Nolet, G. Seismic images of the subducted trailing fragments of the Farallon plate. *Nature* **386**, 266–269 (1997).
- Lithgow-Bertelloni, C. & Richards, M. A. The dynamics of Cenozoic and Mesozoic plate motions. *Rev. Geophys.* **36**, 27–78 (1998).
- Mitrovica, J. X. Haskell [1935] revisited. *J. Geophys. Res.* **101**, 555–569 (1996).
- Hager, B. H. & Richards, M. A. Long-wavelength variations in Earth's geoid: Physical models and dynamical implications. *Phil. Trans. R. Soc. Lond. A* **328**, 309–327 (1989).
- Ricard, Y., Sabadini, R. & Spada, G. Isostatic deformations and polar wander induced by redistribution of mass within the Earth. *J. Geophys. Res.* **97**, 14223–14236 (1992).
- Wasserburg, G. J., MacDonald, G. J. F., Hoyle, F. & Fowler, W. A. Relative contributions of uranium, thorium, and potassium to heat production in the Earth. *Science* **143**, 465–467 (1964).
- Sleep, N. H. Hotspots and mantle plumes: Some phenomenology. *J. Geophys. Res.* **95**, 6715–6736 (1990).
- Davies, G. F. & Richards, M. A. Mantle convection. *J. Geol.* **100**, 151–206 (1992).
- Tackley, P. J. & Stevenson, D. J., Glatzmaier, G. A. & Schubert, G. Effects of multiple phase transitions in a three-dimensional spherical model of convection in Earth's mantle. *J. Geophys. Res.* **99**, 15877–15901 (1994).
- Bercovici, D., Schubert, G. & Glatzmaier, G. A. Three dimensional spherical models of convection in the earth's mantle. *Science* **244**, 893–1016 (1989).
- Megnin, C., Bunge, H.-P., Romanowicz, B. & Richards, M. A. Imaging 3-D spherical convection models: What can seismic tomography tell us about mantle dynamics? *Geophys. Res. Lett.* **24**, 1299–1302 (1997).
- Duffy, T. S. & Ahrens, T. J. Compressional sound velocity, equation of state, and constitutive response of shock-compressed magnesium oxide. *J. Geophys. Res.* **100**, 529–542 (1995).
- Hamilton, W. In *Mesozoic Paleogeography of the Western United States* (eds Howell, D. G. & McDougall, K. A.) 33–70 (Pacific Coast Paleogeography Symp. 2, Pacific Section, Society of Economic Paleontologists and Mineralogists, Los Angeles, 1978).
- Bird, P. Formation of the Rocky Mountains, western United States: a continuum computer model. *Science* **239**, 1501–1507 (1988).
- Jordan, T. E. *et al.* Andean tectonics related to geometry of subducted Nazca plate. *Geol. Study Am. Bull.* **94**, 341–361 (1983).
- Jordan, T. E. & Allmendinger, R. W. The Sierras Pampeanas of Argentina: a modern analogue of rocky mountain foreland deformation. *Am. J. Sci.* **286**, 737–764 (1986).
- Dickenson, W. R. & Snyder, W. S. Plate tectonics of the Laramide orogeny. *GSA Mem.* **151**, 355–366 (1978).

- Livaccari, R. F., Burke, K. & Sengoer, A. M. Was the Laramide orogeny related to subduction of an oceanic plateau. *Nature* **289**, 276–278 (1981).
- Engbreton, D. C., Cox, A. & Thompson, G. A. Correlation of plate motions with continental tectonics: Laramide to Basin and Range. *Tectonics* **3**, 115–119 (1984).
- Zhong, S. & Gurnis, M. Mantle convection with plates and mobile, faulted plated margins. *Science* **267**, 838–843 (1995).
- Tackley, P. J. Effects of strongly temperature-dependent viscosity on time-dependent, 3-dimensional models of mantle convection. *Geophys. Res. Lett.* **20**, 2187–2190 (1993).

Acknowledgements

We thank P. Bird and E. Humphreys for helpful comments on the manuscript.

Correspondence and requests for materials should be addressed to H.-P.B. (e-mail: bunge@princeton.edu).

Functional diversity governs ecosystem response to nutrient enrichment

Florence D. Hulot, Gérard Lacroix, Françoise Lescher-Moutoué & Michel Loreau

Laboratoire d'Ecologie, UMR 7625, Ecole Normale Supérieure, 46 rue d'Ulm, F-75230 Paris Cedex 05, France

The relationship between species diversity and ecosystem functioning is a central topic in ecology today^{1,2}. Classical approaches to studying ecosystem responses to nutrient enrichment have considered linear food chains^{3,4}. To what extent ecosystem structure, that is, the network of species interactions, affects such responses is currently unknown. This severely limits our ability to predict which species or functional groups will benefit or suffer from nutrient enrichment and to understand the underlying mechanisms^{5–8}. Here our approach takes ecosystem complexity into account^{9,10} by considering functional diversity at each trophic level^{11–14}. We conducted a mesocosm experiment to test the effects of nutrient enrichment in a lake ecosystem. We developed a model of intermediate complexity, which separates trophic levels into functional groups according to size and diet. This model successfully predicted the experimental results, whereas linear food-chain models did not. Our model shows the importance of functional diversity and indirect interactions in the response of ecosystems to perturbations, and indicates that new approaches are needed for the management of freshwater ecosystems subject to eutrophication.

We designed our experiment to test the effects of nutrient enrichment on the population density of phytoplankton and zooplankton in the absence and presence of zooplanktivorous fish. The classical approach has been to test qualitative predictions obtained from linear food-chain models. In these models, trophic

Table 1 Qualitative effects of nutrient enrichment as predicted by two linear food-chain models and corresponding experimental results in mesocosms

	Model predictions		Experimental results	
	Prey dependence	Ratio dependence	Without fish	With fish
Carnivores	+	+	–	\$
Herbivores	0	+	ns	ns
Autotrophs	+	+	ns	ns
Phosphorus	0	+	ns	+

Qualitative effects are indicated by their sign: +, 0 and – denote a positive effect, no effect and a negative effect, respectively, of nutrient enrichment on density. Experimental results: + and – denote a significant positive effect and a significant negative effect, respectively ($P \leq 0.05$); brackets, marginally significant effect ($0.05 < P \leq 0.10$); ns, nonsignificant effect ($P > 0.10$); \$, no test possible because the sum of invertebrate carnivores density and fish biomass is senseless.

levels are homogeneous, that is, autotrophs, herbivores and carnivores are described as single compartments. According to the classical prey-dependent model, only the top trophic level and even-numbered trophic levels below the top level should benefit from nutrient enrichment³. Thus, in a three-trophic-level food chain, carnivores control herbivore abundance, and autotrophs, which are released from control by herbivores, control nutrients. Consequently, nutrient enrichment should affect only carnivores and autotrophs (Table 1). In contrast, according to the ratio-dependent model⁴, all trophic levels should benefit from nutrient enrichment (Table 1). Alternative models with either donor control¹⁵ or density-dependent regulation in the top trophic level¹⁶ lead to the same predictions as the ratio-dependent model.

To assess the importance of the internal structure of the pelagic food web^{11–14}, we developed a model of intermediate complexity with several trophic groups per trophic level (Fig. 1, Box 1). As trophic groups are groups of species that share similar prey and similar predators^{7,17}, they represent functional groups from a food-web perspective. Primary producers were divided into edible planktonic algae (A_1), protected planktonic algae (A_2) and periphyton (A_p). We assumed exploitation competition for mineral phosphorus (P) among these groups because of the high nitrogen to phosphorus ratio (20:1). Moreover, periphyton species experienced interference competition, included in the model in the form of a density-dependent loss, because periphyton entirely covered the euphotic zone of the mesocosm walls at the end of the experiment. The second trophic level was divided into small herbivores (H_1), which feed mainly on small edible algae (A_1), and large herbivores (H_2), which feed on both groups of algae (A_1 and A_2). Invertebrate (C_1) and vertebrate (C_2) carnivores are the third trophic level. Invertebrate carnivores (C_1) feed mainly on small herbivores (H_1). Vertebrate carnivores (C_2) are generalist zooplanktivorous fish, which eat small and large herbivores (H_1 , H_2) as well as invertebrate carnivores (C_1). As they feed on more than one trophic level, they are actually omnivorous⁷. Although the density of fish was controlled, their biomass showed large changes depending on treatment because of their strong phenotypic variability. The ratio between final and initial fish biomass varied between 5.9 in the low-nutrient, high-fish treatment to 24.0 in the high-nutrient, low-fish treatment. This requires inclusion of fish biomass as a dynamic variable in the model. We incorporated mutual interference among fish in the form of a density-dependent biomass loss because of the relatively high fish densities in mesocosms.

We used loop analysis¹⁸ to make qualitative predictions about the effects of nutrient enrichment on the equilibrium values of the various functional groups. We then compared these predictions with corresponding experimental results in which trophic groups

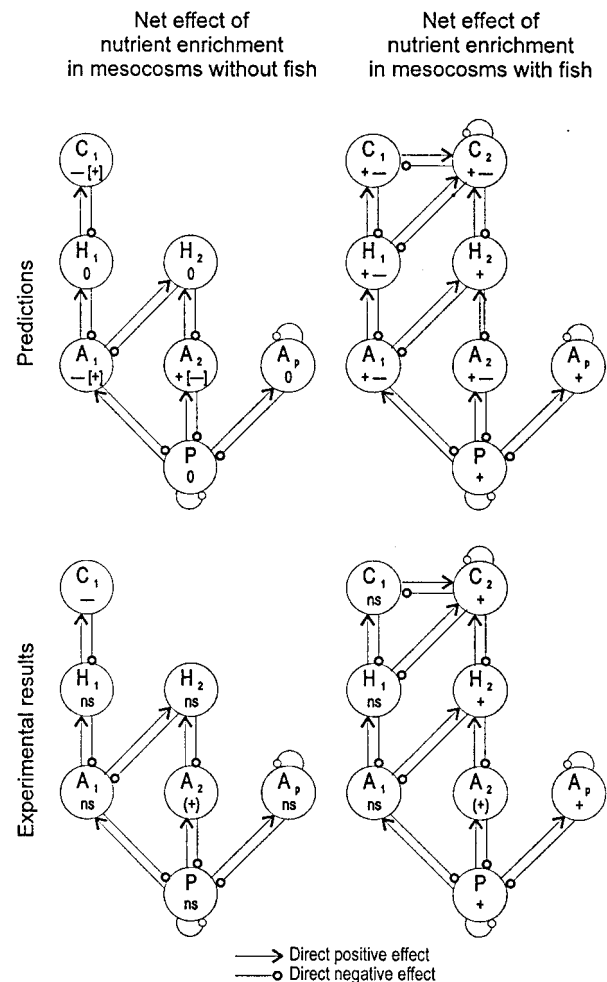


Figure 1 Food-web model describing direct interactions among functional groups in the lake ecosystem, predicted net effects of nutrient enrichment on the equilibrium values of functional groups, and corresponding experimental results in mesocosms. P, mineral phosphorus; A_1 , edible algae; A_2 , protected algae; A_p , periphyton; H_1 , small herbivores; H_2 , large herbivores; C_1 , invertebrate carnivores; C_2 , fish. Model predictions are indicated by their sign: +, 0, – and +– denote a positive effect, no effect, a negative effect and an indetermined effect, respectively, of nutrient enrichment on equilibrium values. Signs in square brackets correspond to possible but unlikely predicted effects (Box 1). Experimental results: + and – denote a significant positive effect and a significant negative effect, respectively ($P \leq 0.05$). Brackets denote a marginally significant effect ($0.05 < P \leq 0.10$); ns denotes a nonsignificant effect ($P > 0.10$).

Table 2 Experimental and statistical results

Group	Without fish		P value	With fish		P value	N	F	N*F
	N ₁	N ₂		N ₁	N ₂				
P	15.1 ± 5.2	20.4 ± 6.8	0.521	7.02 ± 1.02	27.75 ± 3.10	0.0001	0.481	0.09	
A_p	2.5	2.5	0.817	1.1	1.6	0.01	0.5	0.9	
Phytoplankton — mean ± s.e.m. (number of individuals ml ⁻¹)									
A_1	353900 ± 155844	231017 ± 77294	0.536	501120 ± 170503	219633 ± 182364	0.211	0.135	0.275	
A_2	19317 ± 4633	35717 ± 4801	0.067	178625 ± 63167	2057458 ± 1700197	0.098	0.213	0.181	
A	373217 ± 160464	266733 ± 73566	0.650	679745 ± 150701	2377092 ± 1656390	0.401	0.673	0.333	
Zooplankton — mean ± s.e.m. (number of individuals l ⁻¹)									
H_1	1090 ± 277	759 ± 41	0.427	4150 ± 944	3528 ± 814	0.628	0.266	0.417	
H_2	140 ± 31	131 ± 19	0.873	71.8 ± 15.8	178.8 ± 8.8	0.005	0.522	0.373	
H	1230 ± 293	890 ± 43	0.430	4222 ± 957	3707 ± 820	0.717	0.285	0.429	
C_1	51.6 ± 33.3	3.1 ± 1.1	0.024	32.7 ± 17.0	160.5 ± 119.1	0.139	0.293	0.741	
Fish — mean ± s.e.m. (g m ⁻²)									
C_2				11.48 ± 0.64	33.05 ± 2.00	0.0001	0.0036	0.068	

N₁, low nutrient level; N₂, high nutrient level. In mesocosms with fish, means for the two levels of fish are displayed. N, nutrient enrichment effect; F, fish stock effect; N*F, nutrient enrichment*fish stock interaction effect. P values for F and N*F effects are given for information. For all but periphyton (A_p) analyses, some logarithmic transformations were performed to correct for heteroscedasticity.

were distinguished (Fig. 1). As phosphorus and nitrogen were continuously added in mesocosms filled with lake water, our manipulations are press perturbations that should reveal both direct and indirect interactions¹⁹. The duration of the experiment (two months) allowed tens of generations of phytoplankton and few generations of zooplankton, which should be enough to detect effects of direct and indirect interactions^{20,21}. Moreover, the final biomass of zooplanktivorous fish varied between 98 and 390 kg Ha⁻¹, which is comparable to typical fish biomass of mesotrophic to eutrophic lakes²².

Linear food-chain models failed to explain our experimental results. Nutrient enrichment significantly decreased invertebrate carnivore density (C_1) in mesocosms without fish ($P = 0.024$), contradicting the predictions of both the prey-dependent and the ratio-dependent three-level food-chain models (Tables 1 and 2). The effects of nutrient enrichment in mesocosms with fish matched the predictions of the prey-dependent model only for the herbivore level (no significant effect) and the predictions of the ratio-dependent model only for the nutrient level ($P = 0.0001$).

In contrast, the predictions of our intermediate complexity model were consistent with the experimental results. In mesocosms without fish, the experimental results showed a marginally significant increase in the density of protected algae ($P = 0.067$) and a decrease in the density of invertebrate carnivores ($P = 0.024$) with nutrient enrichment in agreement with the model (Table 2). The outcome of a press perturbation such as nutrient enrichment is

determined by the relative strengths of the various propagation paths for the perturbation in the system^{18,23}, and counterintuitive results are common in complex food webs²³. The lower invertebrate carnivore density with nutrient enrichment is counterintuitive but is in agreement with our model predictions (Fig. 1, Box 1). A positive indirect effect of nutrient enrichment is transmitted along the left food chain ($P-A_1-H_1-C_1$) from mineral phosphorus to invertebrate carnivores. But a negative indirect effect is transmitted through consumption of edible algae by large herbivores ($P-A_2-H_2-A_1-H_1-C_1$). This effect is predicted to be stronger because of the greater conversion efficiency of edible algae compared with protected algae (Box 1), leading to a net negative effect. The lack of response of mineral phosphorus concentration, periphyton abundance and densities of small and large herbivores agrees with model predictions.

In mesocosms with fish, nutrient enrichment had positive effects on all levels of the right trophic chain (A_2 : $P = 0.098$; H_2 : $P = 0.005$; C_2 : $P = 0.0001$), as well as on periphyton ($P = 0.01$) and mineral phosphorus ($P = 0.0001$) (Table 2). Note that all the determined model predictions are in agreement with the corresponding experimental results (that is, increases in mineral phosphorus concentration, periphyton abundance and large herbivore density). According to the qualitative analysis of our model, the positive effects of nutrient enrichment on mineral phosphorus, periphyton, large herbivores and fish in mesocosms with fish cannot be explained without incorporating fish predation on invertebrate

Box 1

Mathematical model

The dynamic model corresponding to the graphical model of Fig. 1 is:

$$\frac{dP}{dt} = I - qP - (I_1A_1 + I_2A_2 + I_pA_p)P$$

$$\frac{dA_1}{dt} = A_1(k_1I_1P - m_1 - a_{11}H_1 - a_{12}H_2)$$

$$\frac{dA_2}{dt} = A_2(k_2I_2P - m_2 - a_{22}H_2)$$

$$\frac{dA_p}{dt} = A_p(k_pI_pP - m_p - m'_pA_p)$$

$$\frac{dH_1}{dt} = H_1(b_{11}a_{11}A_1 - d_1 - e_{11}C_1 - e_{12}C_2)$$

$$\frac{dH_2}{dt} = H_2(b_{12}a_{12}A_1 + b_{22}a_{22}A_2 - d_2 - e_{22}C_2)$$

$$\frac{dC_1}{dt} = C_1(f_{11}e_{11}H_1 - \mu_1 - u_{12}C_2)$$

$$\frac{dC_2}{dt} = C_2(f_{12}e_{12}H_1 + f_{22}e_{22}H_2 + v_{12}u_{12}C_1 - \mu_2 - \mu'_2C_2)$$

where I is the external nutrient input; q is the nutrient loss rate; I_j , a_{ij} , e_{ij} and u_{ij} are the consumption rates of P , A_i , H_j and C_j , respectively, by consumer j ; k_j , b_{ij} , f_{ij} and v_{ij} are the conversion efficiencies of consumed food type i into new biomass of consumer j ; m_j , d_j and μ_j are the per capita death rates of A_i , H_j and C_j , respectively; m'_p and μ'_2 are density-dependent interference rates of periphyton and vertebrate carnivores, respectively. The model for mesocosms without fish has $C_2 = 0$.

Loop analysis¹⁸ can be applied directly to the graphical model to predict qualitative changes in equilibrium biomasses following a press perturbation in mineral phosphorus. It was also used to check that the equilibrium is potentially stable, that is, that the first Routh–Hurwitz criteria¹⁸ are met. The system is too complex to check the second Routh–Hurwitz stability criteria analytically, but numerical simulations showed the equilibrium to be stable.

For the model without fish, the sign of the predicted changes is undetermined. However, the sign of these changes can be resolved if additional constraints are taken into account. The equilibrium values for the densities of edible algae, protected algae and invertebrate carnivores can be calculated from the above equations and differentiated with respect to I to evaluate their changes as function of changes in nutrient supply:

$$\frac{\partial A_2^*}{\partial I} = \frac{b_{12}a_{12}}{P^*\alpha}, \quad \frac{\partial A_1^*}{\partial I} = \frac{-b_{22}a_{22}}{P^*\alpha} \quad \text{and} \quad \frac{\partial C_1^*}{\partial I} = \frac{-b_{11}a_{11}b_{22}a_{22}}{P^*\alpha},$$

where $\alpha = b_{12}a_{12}f_{12} - b_{22}a_{22}f_{11}$ and equilibrium values are denoted by an asterisk.

Thus, the qualitative effects of nutrient enrichment on these functional groups hinge on the sign of the lumped parameter α , hence on $(b_{12}a_{12})/b_{22}a_{22}$, the ratio of the direct effects of edible and protected algae on large herbivores, and on I_1/I_2 , the ratio of the nutrient uptakes of edible and protected algae. If $(b_{12}a_{12})/b_{22}a_{22} > I_1/I_2$, nutrient enrichment increases the density of protected algae at the expense of both edible algae and invertebrate carnivores; if the reverse is true, nutrient enrichment increases the biomass of both edible algae and invertebrate carnivores at the expense of protected algae.

As large herbivores are generalists and protected algae escape predation better, the consumption rate of edible algae is expected to be equal to or higher than that of protected algae; hence $a_{12} \geq a_{22}$. The conversion efficiency of protected algae (b_{22}) is generally lower than that of edible algae (b_{12}) for large herbivores¹¹ because of structures that are difficult to digest. Therefore the direct effect of edible algae on the growth rate of large herbivores ($b_{12}a_{12}$) is expected to be stronger than that of protected algae ($b_{22}a_{22}$). Phosphorus uptake per unit of carbon by algae does not appear to depend on size or presence of protections³⁰; hence $I_1 \approx I_2$. Thus we conclude that generally $(b_{12}a_{12})/b_{22}a_{22} > I_1/I_2$, which implies that $\partial A_2^*/\partial I > 0$, $\partial A_1^*/\partial I < 0$ and $\partial C_1^*/\partial I < 0$: in mesocosms without fish, nutrient enrichment is expected to benefit protected algae at the expense of both edible algae and invertebrate carnivores.

For the model with fish, a complex array of factors affects nutrient enrichment effects, which prevents the signs of the net effects from being resolved.

carnivores. Although predation within the level of carnivores may be weak in terms of energy and material flows, it has a major impact on the dynamic response of the community to nutrient enrichment.

Contrary to linear food-chain models, our model of intermediate complexity correctly predicts the qualitative responses of most functional groups to experimental nutrient enrichment. We conducted a sign test on the agreement between all determined predictions (seven in mesocosms without fish and three in mesocosms with fish) and the corresponding experimental results. Only one effect differed (the nutrient enrichment effect on edible algae in mesocosms without fish is not significant whereas we expected a negative effect). Thus, the level of agreement cannot be ascribed to chance ($P \leq 0.025$). Our model is the result of a trade-off between complexity and predictive ability, and, therefore, it does not rule out alternative explanations of experimental results. It does show, however, that taking functional diversity within trophic levels into account is critical to understanding the response of lake ecosystems to environmental perturbations. Some counterintuitive results of nutrient enrichment, such as the decreased density of invertebrate carnivores in the absence of fish, cannot be explained without including multiple functional groups and their resulting indirect interactions. In the prey-dependent linear model, the only indirect interaction that can occur is the trophic cascade¹⁷. In our model, additional indirect interactions, also known in other systems²⁴, are important in the dynamic response of the ecosystems, and yet are ignored in the classical linear food-chain approach²⁵. We conclude that a functional group approach to ecosystems using careful analysis of major species and assumptions on their interactions may provide a better understanding of ecosystem functioning.

These results have important implications for ecosystem management. Lake eutrophication is characterized by increased algal abundance, decreased water transparency and a potential deficit of oxygen²⁶. The classical prey-dependent linear food-chain model predicts that nutrient enrichment should increase algal biomass in lakes with an odd number of trophic levels. Therefore lake restoration usually seeks to eliminate trophic cascades either by direct suppression of zooplanktivorous fish or by addition of piscivorous fish. The mixed success²⁷ of these restorations is understandable in the context of our more complex model. Nutrient enrichment should not affect all algae in the same way. Successful ecosystem restoration needs a new theoretical approach that takes into account community structure and functional diversity. □

Methods

Characterisation of trophic groups

The first trophic level was divided into three functional groups: edible algae (A_1), protected algae (A_2) and periphyton (A_p). Edible algae (A_1) are small (length $< 35 \mu\text{m}$) and unprotected (*Chroomonas* sp., *Coelastrum astroideum*, *C. microporum*, *Colicium* sp., *Cryptomonas* sp., *Cyclotella ocellata*, *Monoraphidium contortum*, *Quadracoccus ellipticus*, *Scenedesmus acuminatus*, *Tetraedron minimum*, *Trachelomonas* sp. and small undetermined unicells). Protected algae (A_2) are large (length $\geq 35 \mu\text{m}$) or protected by thick walls or gelatinous sheaths (*Ceratium hirundinella*, *Cosmarium meneghini*, *Cosmarium* sp., *Crucigenia crucifera*, *C. quadrata*, *C. tetrapedia*, *Dictyosphaerium* sp., *Oocystis lacustris*, *Pediastrum boryanum*, *P. duplex*, *Scenedesmus quadricauda*, *Schroederia indica*, *Staurostrum* sp. and *Synedra ulna*). Herbivores were divided into two functional groups. Small herbivores (H_1), which are roughly 50 to 200 μm long and feed mainly on small edible algae (A_1), include nauplii of cyclopids and calanoids and herbivorous rotifers (*Brachionus angularis*, *B. calyciflorus*, *B. quadridentatus*, *Filinia longiseta*, *Hexarthra mira*, *Keratella cochlearis*, *K. quadrata*, *Lecane* spp., *Polyarthra dolichoptera-vulgaris* and *P. major*). Large herbivores (H_2), which are roughly 400 μm to 2 mm long and feed on the two groups of algae (A_1 and A_2), include cladocerans (*Bosmina*, *Daphnia*, *Ceriodaphnia* and *Diaphanosoma*) and copepodites and adults of calanoid copepods (*Eudiaptomus gracilis*). Invertebrate carnivores (C_1) include carnivorous rotifers (*Asplanchna girodi*, *A. priodonta*) and copepodites and adults of cyclopoid copepods (*Acanthocyclops robustus* and *Thermocyclops crassus*).

Graphical models and qualitative predictions

The nodes in the graphical models correspond to dynamical variables (P: mineral phosphorus; A_1 : edible algae; A_2 : protected algae; A_p : periphyton; H_1 : small herbivores; H_2 : large herbivores; C_1 : invertebrate carnivores; C_2 : fish). Links between nodes correspond to direct effects between trophic groups, determined by the coefficients of the Jacobian

matrix for the dynamical system (Box 1) at equilibrium¹⁸. An arrow or a line ending with a circle from node i to node j corresponds to a positive or a negative direct effect, respectively, of trophic group i on trophic group j . The principle of loop analysis is to identify all propagation pathways of a perturbation from one variable to another and to study feedbacks of variables excluded from these pathways¹⁸.

Experimental mesocosm study

Experimental methods were described in detail elsewhere^{26,28}. Mesocosms were suspended in the lake of Cr teil (suburbs of Paris, France). Each bag contained about 9.5 m³ water (1.5 × 1.5 × 4.3 m deep) and was filled randomly and in several steps between 15 and 20 June 1990 with lake water pumped from 1.5 m depth. A two nutrient level × three fish density level balanced factorial design was implemented with three replicates per treatment. (Two additional nutrient treatments without fish are excluded from our analyses to maintain a balanced design. A fourth level of fish density is also excluded because of fish starvation at the end of the experiment.)

Mesocosms were fertilized with soluble phosphorus (KH_2PO_4) and nitrogen (NH_4NO_3) with a N:P ratio of 20:1 in each treatment three times a week from 27 June to 23 August 1990. The low nutrient treatment was 0.32 $\mu\text{g P l}^{-1} \text{d}^{-1}$ and 6.4 $\mu\text{g N l}^{-1} \text{d}^{-1}$ per enclosure and the high nutrient treatment was 3.16 $\mu\text{g P l}^{-1} \text{d}^{-1}$ and 63.6 $\mu\text{g N l}^{-1} \text{d}^{-1}$ per enclosure.

Two-month-old cyprinids (roach: *Rutilus rutilus*; and common bream: *Abramis brama*) born in the lake were introduced on 29 June 1990. Their mean ($\pm$ s.e.m.) total length and mean wet weight determined from 30 individuals preserved in 4% formalin were 30.0 $\pm$ 0.3 mm and 0.250 $\pm$ 0.009 g, respectively. The three levels of fish density were 0, 10 and 20 individuals per enclosure.

Throughout the experiment, 12 water samples were collected in each enclosure between the surface and the bottom with a 2 l Friedinger bottle and mixed. A 300-ml subsample was preserved in 4% formalin for phytoplankton counts²⁹. The remaining water was filtered through a nylon mesh with a 50- μm aperture and zooplankton was preserved in 4% formalin.

Phytoplankton and zooplankton counts were made on samples taken on 22 August 1990. Phytoplankton counts used the Uterm hl inverted microscope technique at $\times 40$ magnification. The zooplanktonic individuals were counted in Dolfuss chambers under a dissecting microscope. The development of periphyton was monitored weekly using a nominal scale (barren walls, low cover and high cover).

At the end of the experiment (23 August 1990), fish were recaptured, preserved in 10% formalin, dried and then weighed to the nearest milligram.

Statistical analyses

Two separate analyses of variants (ANOVA) tested for effects of nutrient enrichment in mesocosms without fish (one way ANOVA: two levels of nutrient) and effects of nutrient enrichment in mesocosms with fish (two way ANOVA: two levels of nutrient, two levels of fish) on all trophic groups but periphyton. These separate ANOVAs allowed us to compare the experimental results with separate predictions about nutrient effects from our model directly, in the presence or absence of fish. A Mann–Whitney rank test was performed on periphyton abundance. Some logarithmic transformations of density were performed to correct for heteroscedasticity.

Received 7 February; accepted 22 February 2000.

- Johnson, K. H., Vogt, K. A., Clark, H. J., Schmitz, O. J. & Vogt, D. J. Biodiversity and the productivity and stability of ecosystems. *Trends Ecol. Evol.* **11**, 372–377 (1996).
- Chapin III, F. S. *et al.* Biotic control over the functioning of ecosystems. *Science* **277**, 500–504 (1997).
- Oksanen, L., Fretwell, S. D., Arruda, J. & Niemela, P. Exploitation ecosystems in gradients of primary productivity. *Am. Nat.* **118**, 240–261 (1981).
- Arditi, R. & Ginzburg, L. R. Coupling in predator–prey dynamics: Ratio-dependence. *J. Theor. Biol.* **139**, 311–326 (1989).
- Hunter, M. D. & Price, P. W. Playing chutes and ladders: heterogeneity and the relative roles of bottom-up and top-down forces in natural communities. *Ecology* **73**, 724–732 (1992).
- Polis, G. A. Food webs, trophic cascades and community structure. *Aust. J. Ecol.* **19**, 121–136 (1994).
- Persson, L., Bengtsson, J., Menge, B. A. & Power, M. E. In *Food Webs. Integration of Patterns and Dynamics* (eds Polis, G. A. & Winemiller, K. O.) 396–434 (Chapman & Hall, New York, 1996).
- Kaunzinger, C. M. K. & Morin, P. J. Productivity controls food-chain properties in microbial communities. *Nature* **395**, 495–497 (1998).
- McCann, K., Hastings, A. & Huxel, G. R. Weak trophic interactions and the balance of nature. *Nature* **395**, 794–798 (1998).
- Polis, G. A. Stability is woven by complex webs. *Nature* **395**, 744–745 (1998).
- Leibold, M. A. Resource edibility and the effects of predators and productivity on the outcome of trophic interactions. *Am. Nat.* **134**, 922–949 (1989).
- Spiller, D. A. & Schoener, T. W. A terrestrial field experiment showing the impact of eliminating top predators on foliage damage. *Nature* **347**, 469–472 (1990).
- Leibold, M. A. & Wilbur, H. M. Interactions between food-web structure and nutrients on pond organisms. *Nature* **360**, 341–343 (1992).
- Abrams, P. A. Effect of increased productivity on the abundances of trophic levels. *Am. Nat.* **141**, 351–371 (1993).
- Gleeson, S. K. Density dependence is better than ratio dependence. *Ecology* **75**, 1834–1835 (1994).
- Mittelbach, G. G., Osenberg, C. W. & Leibold, M. A. in *Size-structured Populations. Ecology and evolution* (eds Ebenman, B. & Persson, L.) 217–235 (Springer, Berlin, 1988).
- Pimm, S. L., Lawton, J. H. & Cohen, J. E. Food web patterns and their consequences. *Nature* **350**, 669–674 (1991).
- Levins, R. The qualitative analysis of partially specified systems. *Ann. NY Acad. Sci.* **231**, 123–138 (1974).
- Bender, E. A., Case, T. J. & Gilpin, M. E. Perturbation experiments in community ecology: theory and practice. *Ecology* **65**, 1–13 (1984).

20. Leibold, M. A., Chase, J. M., Shurin, J. B. & Downing, A. L. Species turnover and the regulation of trophic structure. *Annu. Rev. Ecol. Syst.* **28**, 467–94 (1997).
21. Menge, B. A. Detection of direct versus indirect effects: were experiments long enough? *Am. Nat.* **149**, 801–823 (1997).
22. Gulati, R. D., Lammens, E. H. R., Meijer, M. -L. & van Donk, E. *Biomanipulation—Tool for Water Management* (Kluwer, Dordrecht, 1990).
23. Yodanis, P. in *Food Webs. Integration of Patterns & Dynamics* (eds Polis, G. A. & Winemiller, K. O.) 192–200 (Chapman & Hall, New York, 1996).
24. Menge, B. A. Indirect effects in marine rocky intertidal interaction webs: patterns and importance. *Ecol. Monogr.* **65**, 21–74 (1995).
25. Power, M. E. Top-down and bottom-up forces in food webs: do plants have primacy? *Ecology* **73**, 733–746 (1992).
26. Lacroix, G. & Lescher-Moutoué, F. Interaction effects of nutrient loading and density of young-of-the-year cyprinids on eutrophication in a shallow lake: an experimental mesocosm study. *Mem. Ist. Ital. Idrobiol.* **48**, 53–74 (1991).
27. DeMelo, R., France, R. & McQueen, D. J. Biomanipulation: hit or myth? *Limnol. Oceanogr.* **29**, 941–948 (1992).
28. Lacroix, G., Lescher-Moutoué, F. & Pourriot, R. in *Aspects of the Genesis and Maintenance of Biological Diversity* (eds Hochberg, M. E., Clobert, J. & Barbault, R.) 162–179 (Oxford Univ. Press, New York, 1996).
29. Hamlaoui, S., Couté, A., Lacroix, G. & Lescher-Moutoué, F. Nutrient and fish effects on the morphology of the Dinoflagellate. *C. R. Acad. Sci. Paris Sci. Vie* **321**, 39–45 (1998).
30. Reynolds, C. S. *Vegetation Processes in the Pelagic: A Model for Ecosystem Theory*. (Ecology Institute, Oldendorf/Luhe, 1997).

Acknowledgements

We thank A. Bertolo, C. de Mazancourt, A. Gonzalez, M. Hochberg, N. Mouquet, O. Petchev, S. Ponsard, M. van Baalen and S. Yachi for comments on the manuscript. This work was supported by grants from the 'Programme Environnement, Vies et Sociétés'.

Correspondence and requests for materials should be addressed to E.D.H. (e-mail: hulot@biologie.ens.fr).

Adaptive plasticity in mate preference linked to differences in reproductive effort

Anna Qvarnström*†, Tomas Pärt†‡ & Ben C. Sheldon*§

* Department of Animal Ecology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18d, SE-752 36 Uppsala, Sweden

‡ Department of Conservation Biology, Swedish University of Agricultural Sciences, Box 7002, SE-750 07 Uppsala, Sweden

§ Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

There is abundant evidence for the existence of marked mate preferences in natural populations, but the occurrence of within-population variation in mate preferences has received little attention^{1–3} and is often regarded as nonadaptive deviation from the optimal norm^{2,3}. Here we show experimentally that the preference of female collared flycatchers *Ficedula albicollis* for male forehead patch size, a sexually selected trait^{4–6}, varies with the time of breeding, an environmental factor with strong effects on reproductive success. Contrary to expectations based on time-constrained choice models^{7,8}, only late-breeding females prefer males with a large patch size. The variation in mate preference matches a seasonal change in female reproductive success: long-term data reveal a positive relationship between female reproductive success and male patch size exclusively in late breeders. In addition, female reproductive effort, as assessed by clutch size, appears to be adjusted relative to both timing of breeding and male phenotype. We conclude that not only can mate preferences display adaptive plasticity within populations, but this plasticity can also be linked to differences in reproductive investment.

During the breeding season male collared flycatchers display a conspicuous white forehead patch which is subject to sexual selection through male–male competition over territories⁴ and sperm competition⁶. Females choose mates in mid-May, soon after their migration back from the African winter quarters to our study area, on Gotland, Sweden. Males assist their females with caring for nestlings, and male parental care is an important determinant of breeding success for females⁹. Breeding success is also strongly seasonally dependent, declining for later breeding birds¹⁰, indicating that females are time-constrained in their choice of mate. Experiments have shown the existence of a trade-off between male investment in mating competition and male investment in parental care¹¹ and suggest that males trade-off these two components of reproductive effort differently depending on their arrival date at the breeding grounds¹². Among males arriving early, large-patched ones (which are more likely to become polygynous⁵ and obtain extra-pair copulations⁶) have a higher premating reproductive effort and are consequently in poorer condition at the time of feeding offspring, as compared with small-patched males; no such pattern is found among late breeders¹².

We investigated how the seasonal increase in large-patched males' allocation of effort to postmating activities affects female reproductive success by analysing a long-term data set collected from the same collared flycatcher population between 1981–1995. The analysis showed that a female's reproductive success is related to both her timing of breeding and the forehead patch size of her mate. The number of offspring fledged from a breeding attempt (when controlling for age and clutch size) was dependent on the interaction between laying date and male forehead patch size (Table 1 and Fig. 1a). Relatively more young were fledged from late breeding attempts when the male had a large forehead patch size as compared with a small forehead patch size, whereas the reverse pattern was found in early breeders (Fig. 1b, c). Thus, the relationship of a female's reproductive success to the patch size of her mate depends on her breeding date. These findings are consistent with a model of

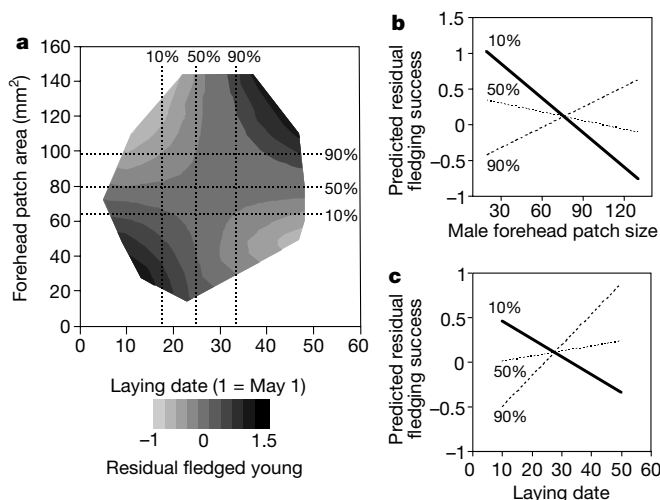


Figure 1 Interaction between time of breeding and male appearance (forehead patch size) on reproductive success of female collared flycatchers. **a**, Contour plot of interaction between laying date and male forehead patch size on residual number of fledged young (controlling for year, parental age effects and clutch size); darker contours represent higher values of reproductive success. Dotted lines with percentages indicate the position of the 10th, 50th and 90th percentiles for the variable whose axis they intersect. **b**, Predicted dependence of residual fledged young on male forehead patch size for three different values of laying date, corresponding to the percentiles marked in **a**. **c**, Predicted dependence of residual fledged young on laying date for three different values of male forehead patch size, corresponding to the percentiles marked in **a**.

† Present address: Department of Biology 0116, University of California at San Diego, La Jolla, California 92093, USA

the evolution of advertisement of parental quality¹³ which predicts a weaker, or even negative, relationship between advertisement and parental investment when the marginal gains from advertisement are increasing, as is likely to be the case for large-patched males breeding early^{6,12}. An alternative explanation, that variation in reproductive success in relation to male patch size is caused by a change in female effort, can be rejected. We found no indication of future fitness costs that such a change in effort should have caused (logistic regression, effect of laying date $\times$ patch area on female survival: $\chi^2 = 0.003$, $n = 1777$, $P = 0.96$; effect of previous laying date $\times$ previous patch area on female's reproductive success the following year: $F_{1,351} = 0.52$, $P = 0.47$).

The seasonal change in the relationship between female reproductive success and male forehead patch size therefore suggests that females should be selected to display seasonal variation in their preference for this trait. We tested whether the female preference for male forehead patch size varied over the course of the breeding season by experimentally manipulating the forehead patch size of unmated males that had been temporarily removed from their territories¹¹. The strength of female preference for a male was measured as the number of days elapsing between a male's release after treatment and his attracting a mate. This experiment revealed that female preference for males with a large forehead patch is seasonally dependent (Table 2). Females choosing from early arrived males showed no preference over controls for males with enlarged forehead patches, but there was a marked female preference for males with enlarged forehead patches in late-arrived birds (Fig. 2). This pattern is in the opposite direction to that predicted by models of mate choice under time constraint, which suggest that mate preferences should become less marked as time is limiting^{7,8}. Instead, the female preference seems to be adjusted according to the seasonal change in how large-patched males allocate their resources between pre- and postmating reproductive activities¹², and thus matches the increased reproductive success of late-breeding females when paired to large-patched males (Fig. 1).

If a male's appearance, in combination with the relative date of the breeding attempt, predicts his postmating reproductive investment, females should be selected to adjust their reproductive effort in response. As with many other single-brooded small passerines breeding in temperate zones¹⁴, the clutch size of female collared

Table 1 Analysis of variation in reproductive success (number of fledged young) of collared flycatchers in relation to age, clutch size, laying date and male forehead patch size SS refers to Sum of Squares

Variable	d.f.	SS	F	P
Year	14	736.13	17.16	< 0.0001
Clutch size	1	312.91	102.12	< 0.0001
Male age	1	4.59	1.50	0.22
Female age	1	7.58	2.47	0.12
Laying date	1	72.25	23.58	< 0.0001
Male forehead patch size	1	38.53	12.57	0.0004
Laying date $\times$ patch size	1	34.97	11.41	0.0007
Error	1,864	5,711.67		

flycatchers declines markedly over the course of the breeding season (Fig. 3a). When the female's mate's forehead patch size is included in the model, however, there is a strong interaction between laying date and male forehead patch size on clutch size (Table 3). The plot of the interaction surface (Fig. 3a) reveals a shape that is consistent with the observed interaction between male forehead patch size and laying date on reproductive success. Whereas clutch size decreases steeply with breeding time in females mated to small-patched males, there is a less marked seasonal decline in clutch size for females paired to large-patched males (Fig. 3b, c). These data suggest that not only do females exhibit adaptive variation in preference over the season, but they also, at least partly, adjust their reproductive effort to match the expected quality of the breeding situation, determined both by timing in the season and the phenotype of their mate.

Why do females not avoid pairing with large-patched males early in the season when such males allocate their main reproductive effort to premating activities? We propose that the answer lies in other benefits associated with choosing large-patched males, which may be independent of breeding time or even show the opposite dependence on breeding time. Comparisons between maternal half-siblings sired by males with different patch sizes have shown that males with larger patches sire offspring that fledge in better condition¹⁵, and male offspring reared early in the year are also more likely to inherit large patches from their fathers¹⁶. Furthermore, large-patched males are likely to provide high quality territories⁵. Thus, a female collared flycatcher choosing a mate faces a complex decision, in which different benefits of choice must be traded off against each other, and in which the relative importance of different benefits changes over time.

Although mate preferences are often described as open-ended, such that extreme expression of secondary sexual traits is always

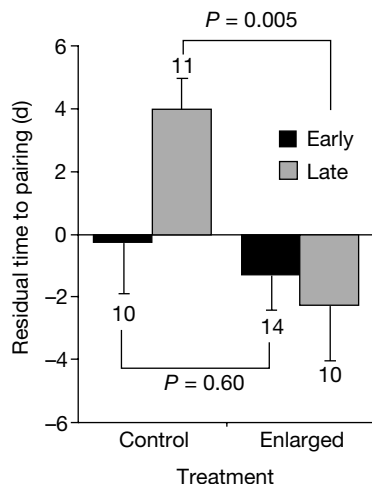


Figure 2 Effect of interaction between experimental manipulation of male forehead patch size and breeding time (early < mean release date after treatment < late) on the speed of mate acquisition, a measure of female preference (mean $\pm$ s.e., controlling for year, age, date of release and original patch size). Numbers above error bars give sample sizes for the groups. In the case of the comparison of early breeding birds, the power to detect an effect of similar magnitude to that in the late breeding birds is $(1 - \beta) = 0.88$.

Table 2 Analysis of variation in female preference (defined as pairing speed) for male collared flycatchers with experimentally manipulated forehead patch size

Variable	d.f.	SS	F	P
Year	1	0.01	0.00	0.98
Male age	1	0.32	0.02	0.90
Date of release*	1	652.74	30.56	0.0001
Original forehead patch size	1	17.97	0.84	0.36
Treatment	1	9.03	0.42	0.52
Release date $\times$ treatment	1	96.07	4.50	0.04
Error	38	811.71		

* Date of release refers to date of release after detention.

Table 3 Analysis of variation in clutch size among collared flycatchers

Variable	d.f.	SS	F	P
Year	14	23.02	3.10	< 0.0001
Male age	1	3.38	6.39	0.01
Female age	1	18.57	35.09	< 0.0001
Laying date	1	21.60	40.78	< 0.0001
Male forehead patch size	1	1.96	3.70	0.05
Laying date $\times$ patch size	1	4.05	7.64	0.006
Error	1,998	1,058.20		

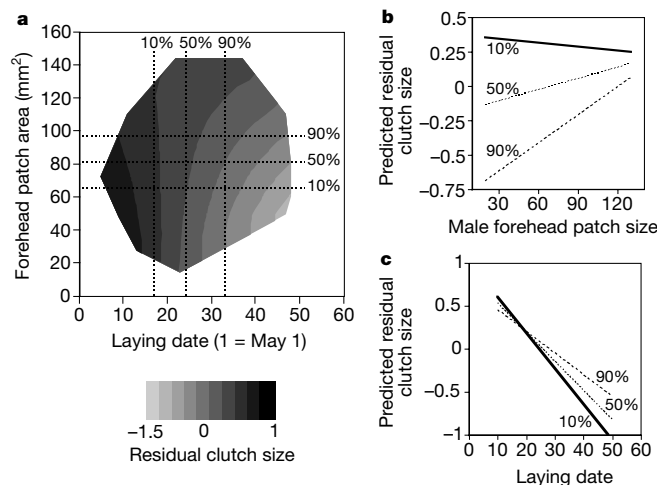


Figure 3 Interaction between timing of breeding and male appearance (forehead patch size) on clutch size laid by female collared flycatchers. **a**, Contour plot of interaction between laying date and male forehead patch size on clutch size (controlling for year and parental age effects); darker contours represent higher values of clutch size. Dotted lines with percentage values indicate the position of the 10th, 50th and 90th percentiles for the

favoured, flexibility in mate choice is, in fact, compatible with most models of sexual selection^{1,2}. In situations where females face trade-offs between different benefits of choice, plasticity in preference is likely to be favoured whenever internal (for example, condition) or external factors affect the pay-offs from choice. Indeed, a recent study on sticklebacks provides evidence for condition-dependent mate choice¹⁷, and the fact that female preference for male throat patch size seems to differ in different populations of house sparrows has been suggested to reflect flexible adjustment to local differences in female requirements¹⁸. Our results show that complex patterns of within-population variation in female preferences may result from short-term environmentally driven changes in expected benefits from choice of males with certain characteristics. Thus, our findings provide direct evidence for adaptive plasticity in mate preferences. □

Methods

Long-term data

Data on reproductive performance (clutch size and number of fledged young) in relation to male phenotype and breeding time were collected using standard methods^{9,19} as part of a long-term study of the collared flycatcher population breeding on southern Gotland (57° 10' N, 18° 20' E) between 1981–1995. Our analyses were restricted to pairs in which both members had been identified as phenotypically pure collared flycatchers, and excluded all cases where brood size or hatching date had been experimentally manipulated²⁰. Fewer than 5% of males are polygynously mated, and we present analyses carried out on all females independently of their mating status, because these give the best impression of the average situation that a female experiences. The results do not change if analyses are done separately on monogamous or either primary or secondary females.

Mate choice experiments

Males were caught when they arrived at the breeding grounds and were randomly assigned to one of two treatment groups. The first group (enlarged) had their forehead patch enlarged by the addition of 4-mm Tipp-Ex correction fluid to the height of the patch. The manipulated size of the forehead patch falls within the natural range. The second group of males (control) were painted with 4-mm Tipp-Ex on their white patch without changing its original size¹¹. All males were ringed and measured using standard methods^{9,19}, and most of them (85%) were removed from their territories for 24 h before they were released in the same area. During detention they were kept outdoors and fed *ad libitum* with mealworms. All nest boxes were checked daily, and a male was defined as chosen by a female as soon as nest material was found in the box that he defended (only females build the nest).

Analysis

Data were analysed using general linear models, in which higher order interactions were sequentially deleted until only those interactions with $P < 0.1$ remained in the model.

variable whose axis they intersect. **b**, Predicted dependence of clutch size on male forehead patch size for three different values of laying date, corresponding to the percentiles marked in **a**. **c**, Predicted dependence of clutch size on laying date for three different values of male forehead patch size, corresponding to the percentiles marked in **a**.

Male and female age were coded as either 'young' (breeding in the year after birth) or 'old' (breeding two or more years after birth); this coding captures most age-related variation in reproductive performance²¹. Significant interactions between continuous variables were visualized by plotting the predicted value of the dependent variable from the final model against the two interacting terms. The contour plots (Figs 1a and 3a) are based on the following equations for the interactions: residual fledged young = $3.30 - 0.116(\text{laying date}) - 0.0420(\text{patch area}) + 0.00152(\text{laying date} \times \text{patch area})$, and residual clutch size = $1.57 - 0.071(\text{laying date}) - 0.0087(\text{patch area}) + 0.00046(\text{laying date} \times \text{patch area})$, respectively.

Received 5 October 1999; accepted 1 March 2000.

- Andersson, M. *Sexual Selection*. (Princeton Univ. Press, Princeton, 1994).
- Jennions, M. D., Petrie, M. Variation in mate choice and mating preferences: a review of causes and consequences. *Biol. Rev.* **72**, 283–327 (1997).
- Widemo, F. & Sæther, S. A. Beauty is in the eye of the beholder: causes and consequences of variation in mating preferences. *Trends Ecol. Evol.* **14**, 26–31 (1999).
- Pärt, T. & Qvarnström, A. Badge size in collared flycatchers predicts the outcome of male competition over territories. *Anim. Behav.* **54**, 893–899 (1997).
- Gustafsson, L., Qvarnström, A. & Sheldon, B. C. Trade-offs between life-history traits and a secondary sexual character in male collared flycatchers. *Nature* **375**, 311–313 (1995).
- Sheldon, B. C. & Ellegren, H. Sexual selection resulting from extra-pair paternity in collared flycatchers. *Anim. Behav.* **57**, 285–298 (1999).
- Real, L. Search theory and mate choice. 1. Models of single-sex discrimination. *Am. Nat.* **136**, 376–405 (1990).
- Gibson, R. M. & Langen, T. A. How do animals choose their mates? *Trends Ecol. Evol.* **11**, 468–470 (1996).
- Gustafsson, L. in *Lifetime Reproduction in Birds* (ed. Newton, I.) 75–88 (Academic, London, 1989).
- Wiggins, D. A., Pärt, T. & Gustafsson, L. Seasonal decline in collared flycatcher *Ficedula albicollis* reproductive success: an experimental approach. *Oikos* **70**, 359–364 (1994).
- Qvarnström, A. Experimentally increased badge size increases male competition and reduces male parental care in the collared flycatcher. *Proc. R. Soc. Lond. B* **264**, 1225–1231 (1997).
- Qvarnström, A. Different reproductive tactics in male collared flycatchers signalled by size of secondary sexual character. *Proc. R. Soc. Lond. B* **266**, 2089–2093 (1999).
- Kokko, H. Should advertising parental care be honest? *Proc. R. Soc. Lond. B* **265**, 1871–1878 (1998).
- Crick, H. Q. P., Gibbons, B. W. & Magrath, R. D. Seasonal changes in clutch size in British birds. *J. Anim. Ecol.* **62**, 263–273 (1993).
- Sheldon, B. C., Merilä, J., Qvarnström, A., Gustafsson, L. & Ellegren, H. Paternal genetic contribution to offspring condition predicted by size of male sexual ornament. *Proc. R. Soc. Lond. B* **264**, 297–302 (1997).
- Qvarnström, A. Genotype-by-environment interactions in the determination of the size of a secondary sexual character in the collared flycatcher (*Ficedula albicollis*). *Evolution* **53**, 1564–1572 (1999).
- Bakker, T. C. M., Künzler, R. & Mazzi, D. Condition-related mate choice in sticklebacks. *Nature* **401**, 234 (1999).
- Griffith, S. C., Owens, I. P. F. & Burke, T. Female choice and annual reproductive success favour less-ornamented male house sparrows. *Proc. R. Soc. Lond. B* **266**, 765–770 (1999).
- Pärt, T. & Gustafsson, L. Breeding dispersal in the collared flycatcher (*Ficedula albicollis*): possible causes and reproductive consequences. *J. Anim. Ecol.* **58**, 305–320 (1989).
- Gustafsson, L. & Sutherland, W. J. The costs of reproduction in the collared flycatcher. *Nature* **335**, 813–815 (1988).

21. Gustafsson, L. & Pärt, T. Acceleration of senescence in the collared flycatcher *Ficedula albicollis* by reproductive costs. *Nature* **347**, 279–281 (1990)

Acknowledgements

We thank I. Ahnesjö, D. Irwin, S. Griffith, L. Gustafsson, C. Hemborg, H. Kokko, T. Price, S. Ulfstrand and F. Widemo for their comments on the manuscript. Financial support was obtained from the Royal Swedish Academy of Sciences, Stiftelsen Zoologisk Forskning (Uppsala University), the Swedish Foundation for International Cooperation in Research and Higher Education (STINT), the Fulbright Commission (A.Q.) the Swedish Natural Sciences Research Council (B.C.S., T.P.) and a Royal Society University Research Fellowship (B.C.S.).

Correspondence and requests for materials should be addressed to A.Q. (e-mail: annaq@biomail.ucsd.edu).

Cross-modal and cross-temporal association in neurons of frontal cortex

Joaquín M. Fuster, Mark Bodner & James K. Kroger*

Neuropsychiatric Institute and Brain Research Institute, UCLA School of Medicine, Los Angeles, California 90095-1759, USA

* Present address: Psychology Department, 3-N-4D Green Hall, Princeton University, Princeton, New Jersey 08544, USA

The prefrontal cortex is essential for the temporal integration of sensory information in behavioural and linguistic sequences^{1,2}. Such information is commonly encoded in more than one sense modality, notably sight and sound. Connections from sensory cortices to the prefrontal cortex support its integrative function^{3–5}. Here we present the first evidence that prefrontal cortex cells associate visual and auditory stimuli across time. We gave monkeys the task of remembering a tone of a certain pitch for 10 s and then choosing the colour associated with it. In this task, prefrontal cortex cells responded selectively to tones, and most of them also responded to colours according to the task rule. Thus, their reaction to a tone was correlated with their subsequent reaction to the associated colour. This correlation faltered in trials ending in behavioural error. We conclude that prefrontal cortex neurons are part of integrative networks that represent behaviourally meaningful cross-modal associations. The orderly and timely activation of neurons in such networks is crucial for the temporal transfer of information in the structuring of behaviour, reasoning and language.

The cardinal function of the association cortex of the frontal lobes, or prefrontal cortex (PFC), is the temporal organization of behaviour (see ref. 2 for a review of the evidence). Prefrontal neurons support this function, in part, by their well documented contribution to short-term memory^{6–10}. Some neurons are attuned to one or more attributes of visual stimuli in memory, such as spatial location or colour^{11,12}. Others are attuned to the associations

of sensory stimuli with motor acts¹³ or with reward¹⁴. Until now, however, there has been no evidence that PFC cells represent the association of sensory items of more than one sensory modality, or that they integrate these items across time. The purpose of our research was to obtain such evidence and establish a role of the PFC in cross-modal and cross-temporal integration. In contrast to previous studies, which have focused on associations between visual stimuli and motor acts, this study demonstrates that prefrontal cells integrate sensory stimuli of different modality that do not share physical dimensions. We show that PFC cells are not only accessible to auditory and visual stimuli, but also treat these stimuli as paired associates in accord with their behavioural role. The paired association by PFC cells takes place across modalities, across time and towards a goal.

The experiment was conducted on two rhesus monkeys trained to perform an audio-visual memory task (Fig. 1a). A trial in the task consisted of the following sequence: (1) a 2-s tone, high-pitch or low-pitch; (2) a 10-s delay; (3) the simultaneous display of two colours, red and green; (4) the manual selection of one colour depending on the tone (red if high tone, green if low tone); and (5) reward if the tone–colour match was correct. During the task, neuronal action potentials (spikes) were recorded extracellularly with microelectrodes from a wide region of dorsolateral frontal cortex bilaterally, including portions of areas 6, 8, 9 and 46 (Fig. 1b).

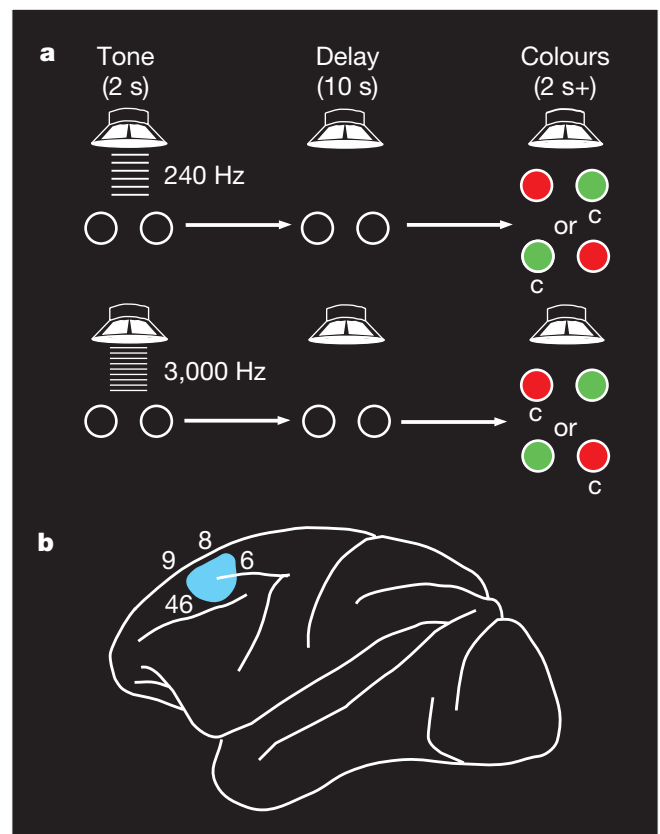


Figure 1 Sound–colour matching task. **a**, Monkey faces panel with two translucent buttons. A trial begins with 240-Hz (low) or 3,000-Hz (high) tone of 2-s duration. Ten-second delay follows, at the end of which buttons are simultaneously lit, red and green. Animal must touch red button if tone was high, green if low. Correct choice (c) is rewarded. Tone and colour position change at random; about 30 s elapse between trials. During inter- and intratrial (delay) intervals, animal rests operant hand on a fixed pedal. Tone is delivered through overhead speaker, with an intensity of 45 dB above background. Each button subtends 8°. Colours are isoluminous (13.5 cd m⁻² ± 0.1 log unit); dominant λ: 620 nm for red, 530 nm for green. **b**, Brain diagram. Numbers indicate frontal cytoarchitectonic areas; blue depicts region containing most of the tone- and colour-differential units recorded.

Table 1 Behavioural coherence of differential firing across trial periods*

	Tone/delay (Δ ₁ versus Δ ₂)	Colours/delay (Δ ₃ versus Δ ₂)	Colours/tone (Δ ₃ versus Δ ₁)	Chosen-colour/tone (Δ ₄ versus Δ ₁)
Coherent	68	66	73	69
Non-coherent	26	28	21	25
Total units	94	94	94	94
	χ ² = 18.76 P < 0.001	χ ² = 15.36 P < 0.001	χ ² = 28.76 P < 0.001	χ ² = 20.59 P < 0.001

* Coherence signifies the same sign of firing difference (delta) in two trial periods paired for comparison.

The firing frequency of isolated neurons was scrutinized for changes occurring in temporal relation to events in the task. In particular, the analysis focused on the cellular response to each tone and each colour. One objective of this analysis was to examine whether the selective reaction of a given cell to a tone was reflected subsequently, at the time of visual choice, by a comparable reaction to the behaviourally associated colour. This relationship would constitute evidence that the cell belongs to the cortical network representing the two associated stimuli in long-term memory (paired cross-modal association).

Most units in the sample of 325 showed changes of firing frequency in temporal relation to one or more task events or periods, although most of these changes were nonspecific for tone or colour. In 94 cells, however, firing at the tone or in the subsequent delay, or both, differed significantly depending on the tone that the monkey had to memorize for correct tone–colour match; the two tones induced a different degree of excitation or inhibition in the cells. Seventy-two of the 94 cells, called hereafter ‘tone-selective cells’, responded with increased firing to one tone and with a lesser excitation, or an inhibition, to the other. (The other 22 differential cells responded to the two tones with different degrees of inhibition.) Whereas in some cells the differential or selective reaction to a tone was limited to the tone period, in others it persisted during part or the entirety of the delay or memory period; and in some cells, tone-selective firing disappeared with the end of the tone only to reappear later in the delay, in anticipation of the colour. Figure 2 illustrates three low-tone selective cells with varying degree and duration of tone-differential firing in the delay period.

Regardless of discharge pattern during the delay, most of the 94 tone-differential units also responded differentially to the two simultaneously presented colours, depending on which of the two was the correct choice. Thus, their responses to colour presentation depended on the tone that had initiated the trial. These colour responses agreed, in most cells, with the behavioural associations of tones and colours. A cell selectively excited by high tone would be more excited at colour presentation in high-tone trials, when red was the correct choice, than in low-tone trials, when green was the correct choice; and vice versa for a low-tone selective cell (Figs 2 and 3). In summary, most differential units showed the same relation of firing to high and low tones as they did to red and green, respectively.

All 94 cells that differentiated the tones—and most of them also the colours—were located in areas 6, 8 or 9/46. Their estimated distribution by area¹⁵ was as follows: 33 cells in area 6; 45 cells in area 8 (mainly 8b, upper prefrontal convexity); and 16 cells in area 9/46. By pooling the stereometric data from both hemispheres and monkeys, most differential units were estimated to lie within a frontal region (Fig. 1, blue area) that straddled and included parts of those three areas. The 231 non-differential units, whether task-related or not, were scattered more or less uniformly throughout the entire dorsolateral frontal region from which cell records were obtained.

Above we have shown in single cells the behavioural–associative–coherence of firing differences (deltas) between high-tone–red and low-tone–green trials across trial periods (tone, delay and colour). Not all cells, however, show coherence across all three periods; some do it only between tone and delay, others between delay and colour, and others between tone and colour. Furthermore, some cells show non-coherent deltas between periods (that is, deltas of opposite sign). The aggregate of the 94 cells, however, shows a highly significant predominance of coherence over non-coherence across all pairs of trial periods (Table 1).

The average deltas between tones and between colours were significantly correlated across trial periods not only in terms of sign but also in terms of magnitude (Fig. 4a). In trials terminating with incorrect choice, the four correlations of average differential firing between trial periods diminished or became negative (Fig. 4b).

By comparing mean spike-frequency responses to the two colours in the chosen-colour period (Δ_4), in high-tone selective cells and in low-tone selective cells, a significant bias for the corresponding colour could be observed in both groups of cells (Fig. 4c).

In conclusion, neurons of the frontal cortex integrate two sense modalities across time in support of behaviour. Because these

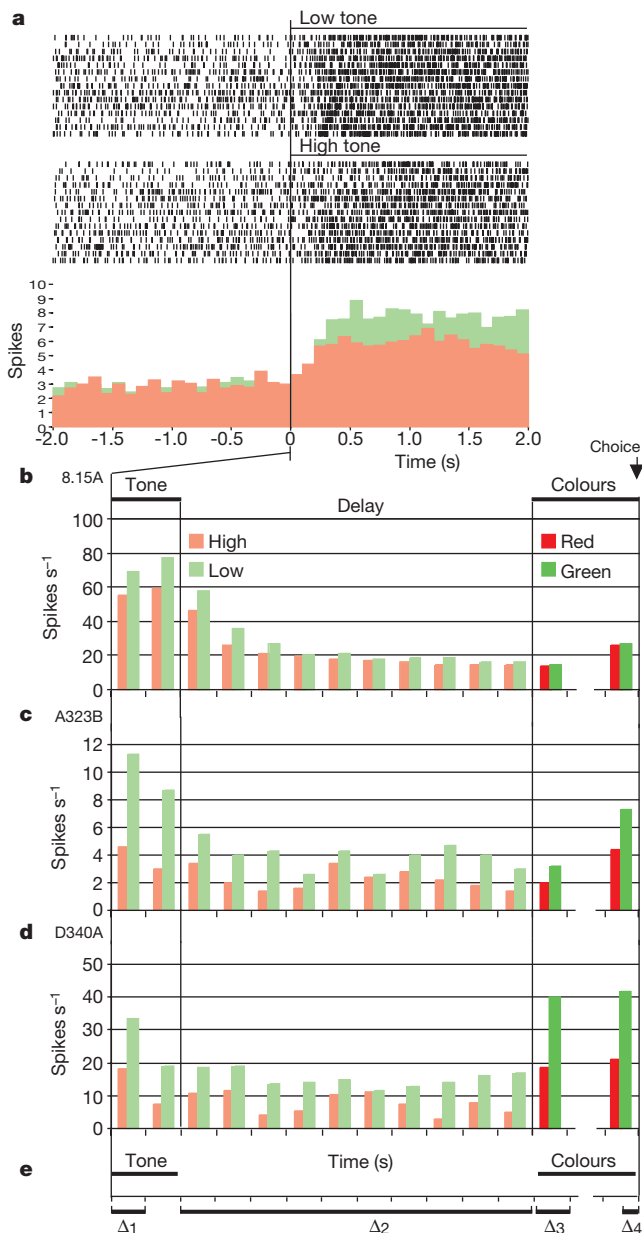


Figure 2 Whole-trial activity of three low-tone selective cells. **a**, Spike rasters and average frequency histograms in baseline and tone period from cell in area 8. Light green indicates low-tone trials; pink indicates high-tone trials. **b**, Frequency histograms through high- and low-tone trials, 1-s time bins in the same cell. (Break of time line indicates variance in reaction time, generally 2.0–3.5 s.) The last bin of the histogram indicates mean firing in 1 s immediately preceding colour choice. **c**, Differential, low-tone selective activity throughout the delay from cell in area 6, bordering 8. As the animal selects a colour (trial's last second), discharge is higher in low-tone trials (green) than in high-tone trials (red). **d**, Increase in low-tone selective activity toward end of delay, anticipating green, from cell near that in **c**. Selectivity is maximal in colour period, when differential firing to colours matches differential firing to tones at trial's start. **e**, Schematic trial sequence indicating four periods in which firing-frequency differences (deltas) were measured for analysis of tone- and colour-dependent firing: Δ_1 , tone; Δ_2 , delay; Δ_3 , colours; Δ_4 , chosen colour.

neurons respond in correlated fashion to behaviourally associated stimuli of separate sensory origin, and because they are found over a broad frontal region, we infer tentatively that they belong to wide cortical networks representing cross-modal associations in permanent storage. Our results deviate from the model of the PFC which posits separate areas for object and spatial representations. The differential cells described here appear to represent cross-modal associations, not specific stimulus properties; furthermore, these cells are 'what' cells and are found in dorsolateral PFC, which has been assumed to represent the 'where' properties of external stimuli¹⁶.

It cannot be concluded, however, that these cells perform only a sensory function—auditory, visual or cross-modal. Their associative properties probably extend to other attributes of the monkey's task that here were not controlled or measured (for example, hand and ocular movements). The representation of non-sensory attributes may be as much a part of the presumptive mnemonic network

to which the cells belong as the sounds and colours of the task. Non-differential reactions of many cells to sounds and colours, as well as the excitatory reactions of differential cells to both tones or both colours below the delta, probably reflect neuronal response to associated sensory qualities that are common to all trials (for example, loudness, brightness, location). Thus, the cells' cross-modal and cross-temporal associations that we have observed would simply contribute, with many others, to the structuring of behaviour by a vast sensory-motor network that extends well beyond our region of interest. Our cells seem to take part in at least three aspects of temporal integration: (1) activation of that network by the tone, which is one of the sensory components of the audio-visual association in permanent storage; (2) sustained low-level activation of the network in the 'working memory' of that association; and (3) reactivation of the network before and during presentation of the associated colour. Different cells appear to participate to a different degree in these three processes. The

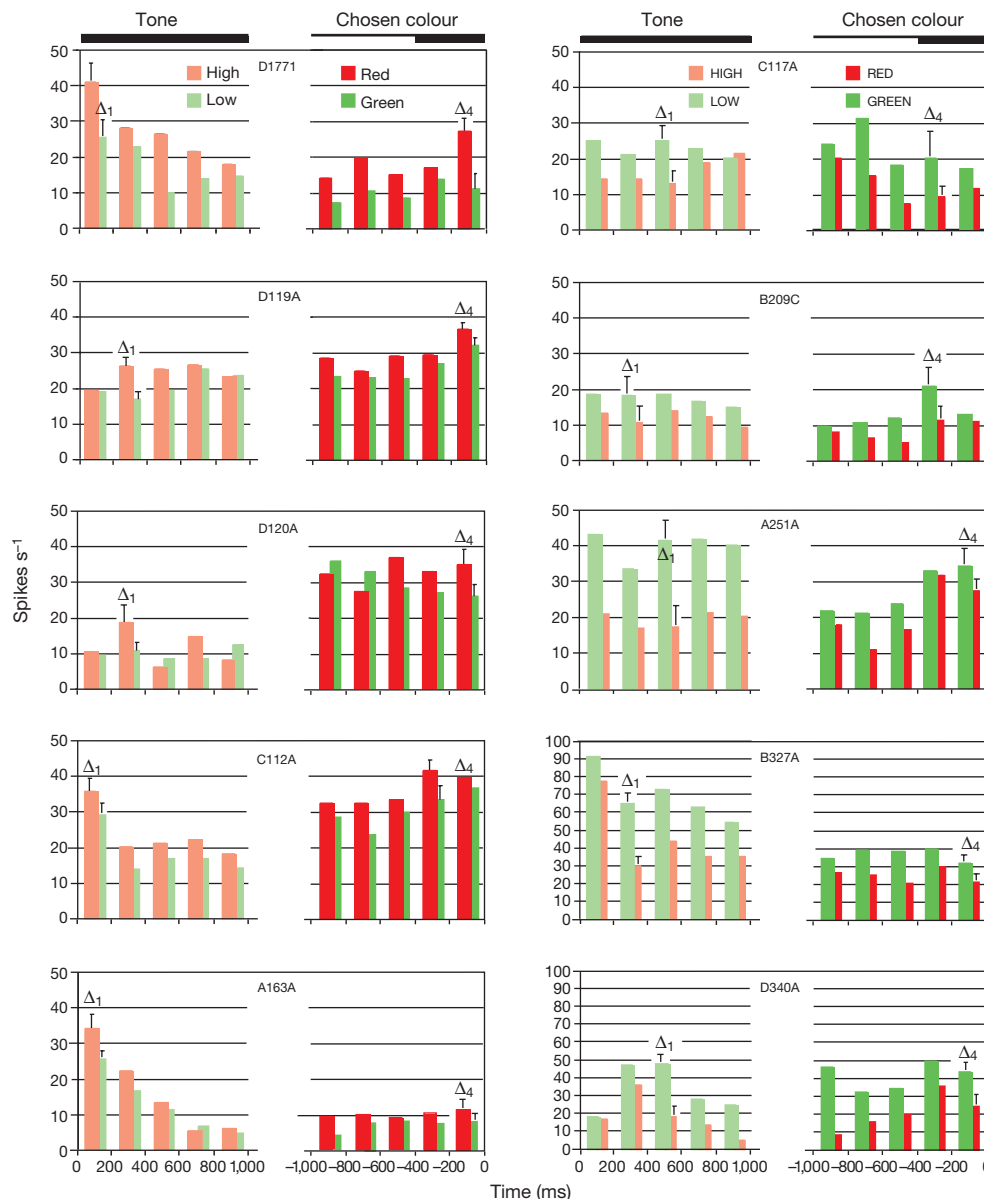


Figure 3 Firing-frequency histograms from ten prefrontal units showing behavioural concordance of firing at tones (first second of tone) and colours (last second of colour). Left, units responding preferentially to high tone and red correct; right, units preferring low tone and green correct. Estimated unit locations: left, from the top, area 9/46, 8b, 8b, 8b

and 8b; right, from the top, area 8b, 9/46, 8b, border 8b/6, and border 8b/6. Note that in all ten cells the firing difference between colours (Δ_4 and adjacent bins) matches in direction the difference between tones some 14 s earlier, at the start of the trial (Δ_1 and adjacent bins).

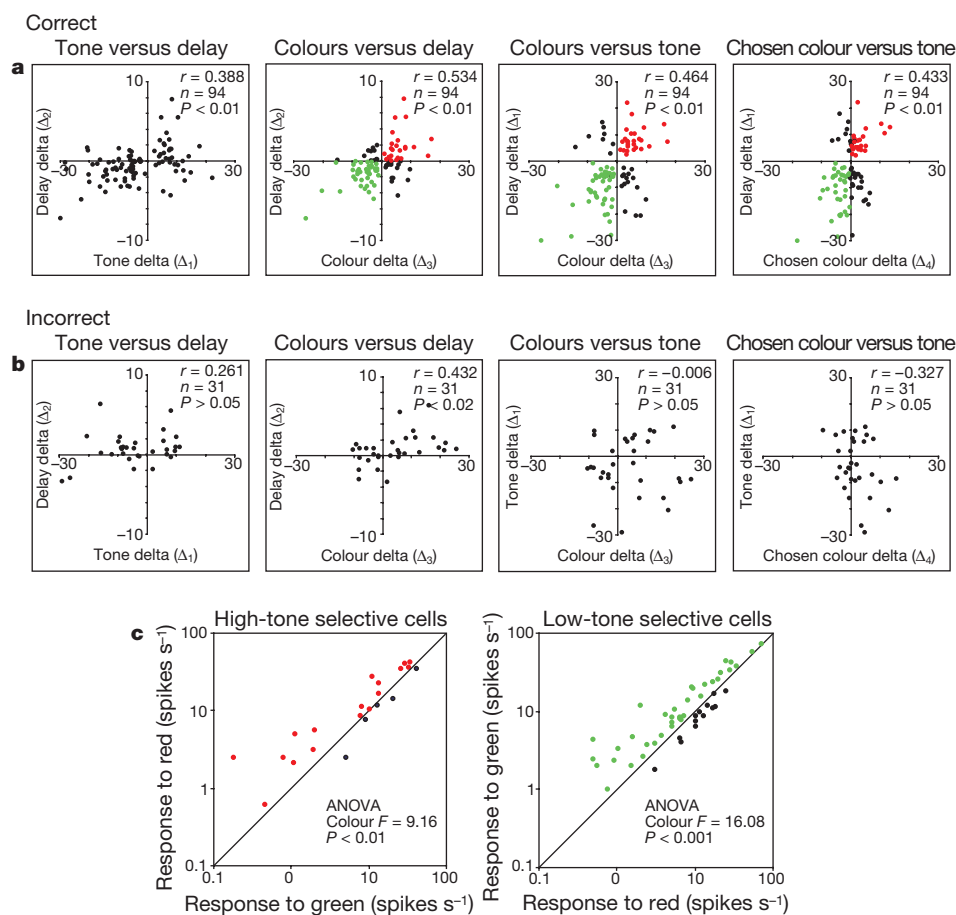


Figure 4 Behaviourally concordant correlations between cell responses to tones and colours. **a**, Correlations between firing deltas across trial periods in 94 tone-differential cells, correct-response trials. **b**, Correlations between deltas in incorrect-response trials.

c, Scatter plots of response to colours by the 72 tone-selective cells (23 high-tone and 49 low-tone). Here, as in **a**, cells with high-tone-red bias are coded red and those with low-tone-green bias are coded green.

continuity of discriminant firing in the cellular aggregate through a trial denotes the selective and orderly activation of the network in the temporal transfer of cross-modal information, from the tone to the monkey's hand.

Beyond sensory cortices, complex sensory features are integrated in associative areas of temporal and parietal cortex^{17–20}. Some of these areas are involved in intermodal association^{21–24}. The integration of sensory stimuli with motor acts engages these areas in dynamic interaction with frontal cortex^{25–27}, especially the PFC when the integration is across time, as our data indicate. The frontal region that we have explored is one of the highest stages of sensory-motor integration. It receives afferent connections from multiple areas of sensory association^{3–5,28}. Our data indicate that a subregion within it (Fig. 1b, blue area) is especially rich in audiovisual associative cells. Accordingly, lesions of it produce deficits in conditional auditory and visual discrimination²⁹. That frontal region is also strategically situated for the organization of action in time. The available evidence² supports a critical role of dorso-lateral prefrontal and premotor cortices in motor set and motor sequences, and thus, more broadly, in the temporal organization of behaviour. This role extends to the organization of language, which depends on the temporal integration of stimuli (words) encoded in the two sense modalities of our cross-modal task. □

Methods

Behaviour

The animals were trained in the audiovisual task to a criterion of a correct performance of 75% or higher. Training was laborious and took several months, mainly because of the

difficulty that monkeys have in learning auditory discriminations. Even after reaching criterion, performance remained variable throughout the experiment. No attempt was made to restrict or control the animals' ocular motility. As verified by a previous study³⁰, monkeys performing colour choices, as they do here at the end of each trial, foveate the colour they are about to choose by hand. The experiment was conducted by strict adherence to animal use guidelines of the UCLA School of Medicine, the National Institutes of Health and the Society for Neuroscience.

Recording

The database for the present study consists of 325 units. Each cell was selected according to the following criteria: (1) clear separation of its action potentials from those of other cells; (2) stability of intertrial baseline firing; (3) records from a sufficient number of trials with each of the test stimuli; and (4) monkey's performance was at least 55% correct. The number of trials analysed varied from cell to cell depending on cell stability, firing frequency and the monkey's performance. On average, the 325 cells that met the criteria were recorded through a minimum of 10 correct-response trials with each tone. The median level of performance during recording was 76.6% correct. Recording sites were estimated by histology in one animal, and by stereotaxia and cortical landmarks in the other.

Data analysis

In a first stage of analysis, trial-related changes of firing frequency were computed for the 325 units in terms of deviations from baseline firing, that is, from the average firing of each cell in the period (10 s or longer) immediately preceding the trial. Spike discharge in the course of the trial was divided in time bins of 0.2, 0.5, 1.0, 5.0 or 10 s (last two bin-durations applied only to delay period). Average per bin difference from baseline, in spikes s^{-1} , was calculated across all correct-response trials with each tone. Average differences between blocks of trials with different tone were submitted to a Student's *t*-test for correlated means, with intertrial variance in the error term (significance of differences ascribed to $P < 0.05$).

The analysis of deviations from baseline firing led to the identification of the 94 units on which the cross-modal and cross-temporal findings of this study are based. They were

selected by the following criteria: (1) stable intertrial baseline, without statistically significant differences in average firing rate between pre-low-tone and pre-high-tone baselines (the overall average baseline firing for the 94 cells was 10.03 spike s⁻¹; and (2) statistically significant differences in firing between the two tones, at tone presentation or in the subsequent delay. These cells were submitted to the correlational analysis of average tone- and colour-related differences (deltas) between trial periods. Four such periods were the object of analysis: (1) first 1 s of tone; (2) 10-s delay; (3) first 1 s of colour presentation; and (4) 400-ms period immediately preceding manual choice of colour. Tone and colour periods (1, 3 and 4) were divided in 200-ms bins, and the entire delay period was treated as a 10-s bin. Four tone- or colour-related differences or deltas (Fig. 2e) were determined between high-tone (red choice) trials and low-tone (green choice) trials: Δ_1 (tone), largest 1-bin difference between responses to the two tones; Δ_2 (delay), tone-related difference in delay firing; Δ_3 (colours), largest 1-bin difference in response to the two simultaneous colours; and Δ_4 (chosen colour), largest 1-bin difference in pre-choice firing. Table 1 tabulates cells by correlation of delta sign across trial periods. Figure 4a, b shows delta correlations by magnitude. Expected correlations by chance were calculated by shuffling 1,000 times the delta values from both animals (94 differential units) in correct-response trials: the mean coefficient (r) thus obtained for all four interperiod correlations (Δ_1 versus Δ_2 , Δ_3 versus Δ_2 , Δ_3 versus Δ_1 , and Δ_4 versus Δ_1) ranged between minus 0.002 and plus 0.003 (s.d. 0.103–0.105).

Received 15 February; accepted 29 March 2000.

- Luria, A. R. *Higher Cortical Functions in Man* (Basic Books, New York, 1966).
- Fuster, J. M. *The Prefrontal Cortex* 3rd edn (Lippincott–Raven, Philadelphia, 1997).
- Jones, E. G. & Powell, T. P. S. An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. *Brain* **93**, 793–820 (1970).
- Pandya, D. N. & Yeterian, E. H. In *Cerebral Cortex* (eds Peters, A. & Jones, E. G.) 3–61 (Plenum, New York, 1985).
- Romanski, L. M. *et al.* Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nature Neurosci.* **2**, 1131–1136 (1999).
- Fuster, J. M. Unit activity in prefrontal cortex during delayed-response performance: Neuronal correlates of transient memory. *J. Neurophysiol.* **36**, 61–78 (1973).
- Niki, H. Differential activity of prefrontal units during right and left delayed response trials. *Brain Res.* **70**, 346–349 (1974).
- Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* **61**, 331–349 (1989).
- Bodner, M., Kroger, J. & Fuster, J. M. Auditory memory cells in dorsolateral prefrontal cortex. *NeuroReport* **7**, 1905–1908 (1996).
- Romo, R., Brody, C. D., Hernández, A. & Lemus, L. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* **399**, 470–473 (1999).
- Fuster, J. M., Bauer, R. H. & Jervey, J. P. Cellular discharge in the dorsolateral prefrontal cortex of the monkey in cognitive tasks. *Exp. Neurol.* **77**, 679–694 (1982).
- Rao, S. C., Rainer, G. & Miller, E. K. Integration of what and where in primate prefrontal cortex. *Science*, **276**, 821–824 (1997).
- Rainer, G., Asaad, W. F. & Miller, E. K. Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature* **393**, 577–579 (1998).
- Watanabe, M. Reward expectancy in primate prefrontal neurons. *Nature* **382**, 629–632 (1996).
- Petrides, M. & Pandya, D. N. In *Handbook of Neuropsychology* (eds Boller, F. & Grafman, J.) 17–58 (Elsevier, Amsterdam, 1994).
- Wilson, F. A. W., Scalaidhe, S. P. O. & Goldman-Rakic, P. S. Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science* **260**, 1955–1958 (1993).
- Desimone, R., Albright, T. D., Gross, C. G. & Bruce, C. Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* **4**, 2051–2062 (1984).
- Miyashita, Y. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* **335**, 817–820 (1988).
- Iwamura, Y., Tanaka, M., Sakamoto, M. & Hikosaka, O. Rostrocaudal gradients in the neuronal receptive field complexity in the finger region of the alert monkey's postcentral gyrus. *Exp. Brain Res.* **92**, 360–368 (1993).
- Tanaka, K. Neuronal mechanisms of object recognition. *Science* **262**, 685–688 (1993).
- Colombo, M. & Gross, C. G. Responses of inferior temporal cortex and hippocampus neurons during delayed matching to sample in monkeys (*Macaca fascicularis*). *Behav. Neurosci.* **108**, 443–455 (1994).
- Gibson, J. R. & Maunsell, J. H. R. Sensory modality specificity of neural activity related to memory in visual cortex. *J. Neurophysiol.* **78**, 1263–1275 (1997).
- Hikosaka, K., Iwai, E., Saito, H. & Tanaka, K. Polysensory properties of neurons in the anterior bank of the caudal superior temporal sulcus of the macaque monkey. *J. Neurophysiol.* **60**, 1615–1637 (1988).
- Zhou, Y. & Fuster, J. M. Neuronal activity of somatosensory cortex in a cross-modal (visuo-haptic) memory task. *Exp. Brain Res.* **116**, 551–555 (1997).
- Fuster, J. M., Bauer, R. H. & Jervey, J. P. Functional interactions between inferotemporal and prefrontal cortex in a cognitive task. *Brain Res.* **330**, 299–307 (1985).
- Wise, S. P., Boussaoud, D., Johnson, P. B. & Caminiti, R. Premotor and parietal cortex: corticocortical connectivity and combinatorial computations. *Annu. Rev. Neurosci.* **20**, 25–42 (1997).
- Tomita, H., Ohbayashi, M., Nakahara, K., Hasegawa, I. & Miyashita, M. Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature* **401**, 699–703 (1999).
- Matelli, M., Luppino, G. & Rizzolatti, G. Patterns of cytochrome oxidase activity in the frontal agranular cortex of macaque monkey. *Behav. Brain Res.* **18**, 125–137 (1985).
- Petrides, M. The effect of periacute lesions in the monkey on the performance of symmetrically and asymmetrically reinforced visual and auditory go, no-go tasks. *J. Neurosci.* **6**, 2054–2063 (1986).
- Fuster, J. M. & Jervey, J. P. Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task. *J. Neurosci.* **2**, 361–375 (1982).

Acknowledgements

We thank W. Bergerson and B. Lubell for technical assistance; L. Fairbanks for statistical

counsel; and C. R. Gallistel, J. Maunsell, W. T. Newsome, T. Sejnowski and Y. Zhou for valuable comments on the manuscript. This work was supported by grants from the National Institute of Mental Health.

Correspondence and requests for materials should be addressed to J.M.F. (e-mail: joaquin@ucla.edu).

Phenotypic suppression of *empty spiracles* is prevented by *buttonhead*

Frieder Schöck[†], Joachim Reischl[‡], Ernst Wimmer[‡], Heike Taubert^{*}, Beverly A. Purnell[†] & Herbert Jäckle^{*}

^{*} Max-Planck-Institut für Biophysikalische Chemie, Abt. Molekulare Entwicklungsbiologie, Am Fassberg 11, 37077 Göttingen, Germany

[‡] Lehrstuhl für Genetik, Universität Bayreuth, Universitätsstr. 30, NW1, 95440 Bayreuth, Germany

Unlike the trunk segments, the anterior head segments of *Drosophila* are formed in the absence of pair-rule^{1,2} and HOX-cluster gene³ expression, by the activities of the gap-like genes *orthodenticle* (*otd*), *empty spiracles* (*ems*) and *buttonhead* (*btd*)^{4,5}. The products of these genes are transcription factors^{6,7}, but only EMS has a HOX-like homeodomain^{8,9}. Indeed, *ems* can confer identity to trunk segments¹⁰ when other HOX-cluster gene activities are absent^{3,11}. In trunk segments of wild-type embryos, however, *ems* activity is prevented by phenotypic suppression¹⁰, in which more posterior HOX-cluster genes inactivate the more anterior without affecting transcription or translation¹². *ems* is suppressed by all other Hox-cluster genes and so is placed at the bottom of their hierarchy¹⁰. Here we show that misexpression of EMS in the head transforms segment identity in a *btd*-dependent manner, that misexpression of *BTD* in the trunk causes *ems*-dependent structures to develop, and that EMS and *BTD* interact *in vitro*. The data indicate that this interaction may allow *ems* to escape from the bottom of the HOX-cluster gene hierarchy and cause a dominant switch of homeotic prevalence in the anterior–posterior direction.

Combined activities of *otd*, *ems* and *btd*⁵ generate and specify *Drosophila* head segments (Fig. 1a–d) in the absence of pair-rule and homeotic gene activities⁴ (Fig. 1e). *btd* alone is required for development of the mandibular segment, *btd* plus *ems* for the intercalary segment, and *btd*, *ems* plus *otd* for the antennal segment (Fig. 1e). Misexpression of *btd* or *otd* in the prospective head region failed to cause homeotic transformations^{13,14} showing that neither of the two genes carries the proposed homeotic function in head segmentation⁵. To explore the untested homeotic role of *ems* and to address a possible cooperation with *btd*, we misexpressed the *ems* protein (EMS) in the *btd* domain of otherwise wild-type embryos. EMS expression was achieved by an *ems* complementary DNA transgene under control of the *btd* cis-acting promoter region¹³.

EMS expression in the *btd* domain of wild-type embryos caused a second intercalary-like *engrailed* expression domain in place of the mandibular segment (Fig. 2a and d). Furthermore, these embryos developed a duplicate set of intercalary cuticle elements in place of mandibular structures (compare Fig. 2b, c with e, f). We observed the same results in response to EMS expression in the anterior third of blastoderm embryos mediated by a Gal4/UAS system¹⁵. However, misexpression of OTD in the *btd* domain had no effect on head

[†] Present addresses: Department of Genetics, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115, USA (F.S.); American Association for the Advancement of Science, 1200 New York Avenue NW, Washington DC 20005, USA (B.A.P.).

formation (data not shown). Thus, among the three head gap-like genes only *ems* carries both early segmentation and homeotic selector gene function. However, *ems* misexpression in several head segments caused only the mandibular into intercalary segment transformation. As the intercalary segment also depends on *btd*, and *ems* did not have transforming activity in *btd* mutant embryos (data not shown), we conclude that *ems* activity is able to specify head segment identity only when acting in concert with *btd*. Notably, the direction of the *ems*-dependent transformation is from a posterior into a more anterior segment identity. This is opposite to the direction of transformation in response to ectopically expressed HOX-cluster genes in the trunk (reviewed in ref. 12).

BTD is a transcriptional activator with *in vitro* properties that are indistinguishable from those of human Sp1 (refs 7,16). However, whereas transgene-derived *btd* activity causes a full rescue of all head segments that are deleted in *btd* mutant embryos^{7,16}, Sp1 activity could rescue only mandibular segment development. Similarly, expression of the fusion protein VP16^{BTDzf}, which contains the VP16 transactivator region¹⁷ fused to BTD's zinc finger domain⁷, only rescued mandibular development (Fig. 3a; Fig. 2g–i). Conversely, expression of BTD^{Sp1zf}, in which the BTD zinc finger domain was replaced by the zinc finger domain of human Sp1, mediated a complete rescue of *btd* mutant embryos (Fig. 3a). Thus, BTD must contain specific features outside its zinc finger domain that are needed for intercalary segment development.

To identify the BTD region necessary for the *ems*-dependent intercalary development, we then asked whether BTD can physically interact with EMS *in vitro* and which parts of BTD are involved. BTD is able to bind [³⁵S]methionine-labelled EMS *in vitro* (Fig. 3b).

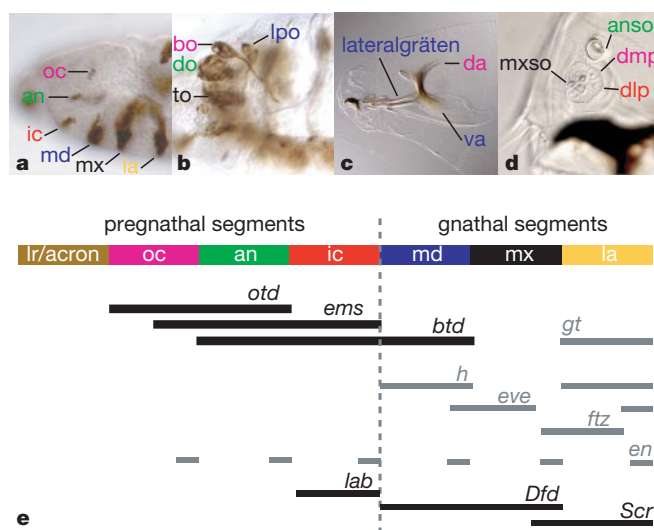


Figure 1 Representation of the *Drosophila* head. **a, b**, Heads of wild-type embryos showing segmental stripes of Engrailed (**a**) and larval head sensory organs stained with 22C10 monoclonal antibody (**b**). **c, d**, Cuticle head markers of larvae (**c**) including the antennomaxillary complex (enlarged; **d**). Segmental origin is labelled according to the colour code shown in **e**. Note the Bolwig organ (bo), the dorsal arms (da) and the dorsomedial papilla (dmp) (ocular segment), the antennal sense organ (anso) and the dorsal organ (do) (antennal segment), the dorsolateral papilla (dlp) (intercalary segment), the lateropharyngeal organ (lpo) and ventral arms (va) (mandibular segment), the maxillary sense organ (mxso) and the terminal organ (to) (maxillary segment). For details see ref. 26. **e**, The head anlage showing the position and colour code for labial (la), maxillary (mx), mandibular (md), intercalary (ic), antennal (an), ocular (oc) segments and the labrum (lr) as part of the acron²⁷. Black bars, blastoderm expression domains of *orthodenticle* (*otd*), *empty spiracles* (*ems*), *buttonhead* (*btd*) and the homeotic genes *labial* (*lab*), *Deformed* (*Dfd*) and *Sex combs reduced* (*Scr*); grey bars, expression domains of *giant* (*gt*), the pair-rule genes *hairy* (*h*), *even-skipped* (*eve*) and *fushi tarazu* (*ftz*) and the segment polarity gene *engrailed* (*en*).

This interaction involves the amino-terminal region of the protein (Fig. 3c and d). We could not define a specific domain as several parts of the N-terminal region interacted with EMS (Fig. 3e and f), excluding the zinc finger domain (not shown). Sp1, which has the same biochemical features as BTD¹⁶, failed to interact with EMS (data not shown). The yeast two-hybrid system also showed a direct interaction between EMS and BTD's N-terminal region that does not involve the homeodomain of EMS (Fig. 3g and h).

Next we examined whether the BTD mutants that interact with EMS were sufficient to allow homeotic EMS activity *in vivo*. We performed transgene-dependent rescue experiments in which BTD deletion mutants were expressed in *btd* mutant embryos. N-BTD, a protein composed of the combined DNA-binding and N-terminal region, rescues all head segments of *btd* mutant embryos, whereas C-BTD, a protein lacking the N terminus, causes mandibular segment development in all *btd* mutant embryos but restores only partial intercalary development in rare cases (<5%; *n* = 66 embryos; Fig. 3a). Furthermore, BTD variants that lack various portions of the N terminus were able to restore intercalary segment development fully, indicating that BTD-dependent intercalary development depends on the parts of its N-terminal region that also allow physical interaction with EMS.

To investigate whether the BTD and EMS interaction causes homeotic transformations in other parts of the embryo, we made use of the observation that *ems* is also expressed in the trunk region of the embryo from stage 9 onwards^{8,9}. Lack of *ems* activity causes no alteration in trunk segments except that the 'filzkörper', a morpho-

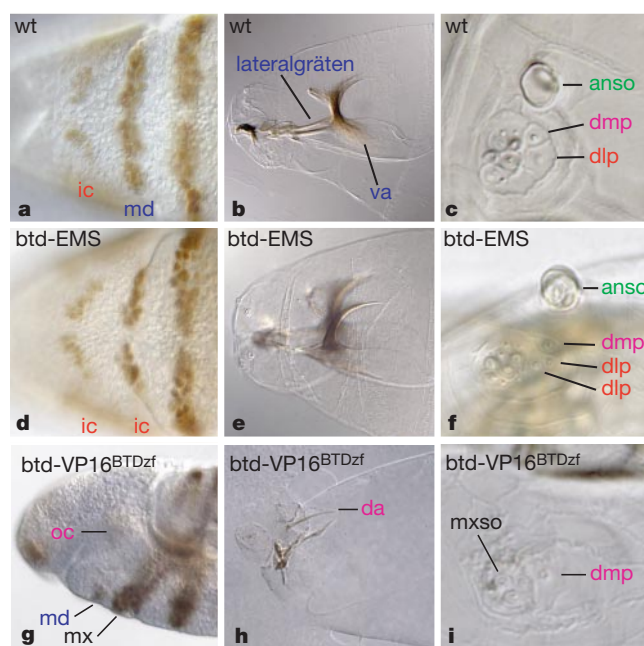


Figure 2 Misexpression of *ems* in the *btd* domain of wild-type embryos transforms mandibular to intercalary identity. **a–c**, Wild-type embryos. **a**, Mandibular (md) and intercalary (ic) Engrailed stripes (ventral view). **b**, Head skeleton. **c**, Antennomaxillary complex with a single dorsolateral papilla (dlp) of intercalary origin. **d–f**, Corresponding embryos containing two copies of the *btd-EMS* construct. **d**, The two most anterior Engrailed stripes showing the characteristics of the intercalary Engrailed stripe. **e**, Mandibular skeleton structures (ventral arms and lateralgräten)⁷ are reduced and absent, respectively. **f**, Duplicated dlp indicates intercalary structures at the expense of mandibular development. Controls confirmed that OTD does not cause transformations when acting with EMS and/or BTD in the intercalary segment anlage^{13,14}. **g–i**, Transgene-dependent rescue of *btd*-dependent mandibular segment by VP16^{BTDzf} expression in *btd* mutants (see text) is similar to the Sp1-dependent rescue¹³. Note a reduced mandibular Engrailed stripe (**g**), the head skeleton including dorsal arms (da; **h**) and the antennomaxillary complex lacking dlp and anso (**i**). For comparison see Fig. 1a–d.

logically distinct structure of the last abdominal segment¹⁸, fails to develop. In the absence of all HOX-cluster gene activities, however, the trunk segments alter identity and develop *ems*-dependent sclerotic head plates¹⁰. Their formation can be phenotypically suppressed by the co-expression of any gene of the HOX-cluster including *labial*, *Deformed* or *Sex combs* reduced¹⁰, which are normally expressed and required in the cephalic region of the embryo (Fig. 1e; reviewed in ref. 3). Therefore, *ems* may be a disconnected member of the ancient HOX-cluster^{3,11}, acting at the bottom of the functional HOX-cluster hierarchy¹⁰.

The *ems*-dependent homeotic transformation in the head region may be because of a requirement of EMS to cooperate with BTd to escape from phenotypic suppression. To test this in the trunk region, we expressed BTd from a heat-shock-inducible transgene at various early embryonic stages. Ectopic BTd expression up to and during blastoderm stage had no effect on trunk segmentation (data not shown). However, BTd expression during stage 7–9 of embryogenesis, when *ems* is initially expressed in the prospective trunk region^{8,9}, caused a range of phenotypes. These included the development of sclerotic head plates reminiscent of the *ems*-depen-

dent structures observed in embryos that lack the HOX-cluster genes¹⁰.

Most BTd misexpressing embryos develop fusions of trunk segments to varying degrees (149 cases out of 218 heat-shocked embryos examined; compare Fig. 4a and b). However, such segment fusions were also observed at a similar frequency in BTd-expressing homozygous *ems* mutant embryos (32 of 54 embryos examined). Thus, ectopic BTd activity causes metamorphosis defects independent of *ems* activity. However, BTd-expressing wild-type embryos also develop sclerotic plates (68 cases out of 218 heat-shocked embryos examined; Fig. 4b). In a few cases (11 embryos), segmentation was completely abolished (Fig. 4c), and sclerotic plates were found (Fig. 4c and d). Sclerotic plates were never observed in embryos lacking *ems* activity (16 embryos examined). Thus, their formation in the trunk region of embryos depends on combined BTd and EMS activities. *ems* escapes phenotypic suppression by the HOX-cluster genes without BTd affecting the HOX-cluster gene transcription or translation (Fig. 4e).

The results provide evidence that combined BTd and EMS activities specify the intercalary head segment identity. The gnatho-cephalic homeotic genes *labial* and *Deformed* are normally expressed in intercalary and mandibular head segments, respectively (Fig. 1e; reviewed in ref. 10), and their products cause phenotypic suppression of EMS. As the intercalary segment development is dependent on the regions of BTd which can associate with EMS *in vitro*, it is probable that the EMS–BTd interaction releases the phenotypic suppression. EMS can also overcome phenotypic suppression by the HOX-cluster genes in the trunk when co-expressed with ectopic BTd. We propose that the interaction with BTd allows EMS to relocate from the bottom to the top of the HOX-cluster gene hierarchy. EMS then functions in an anterior-prevalent manner, that is, in the opposite direction to other HOX-cluster genes. The unique homeotic feature of *ems* among the head

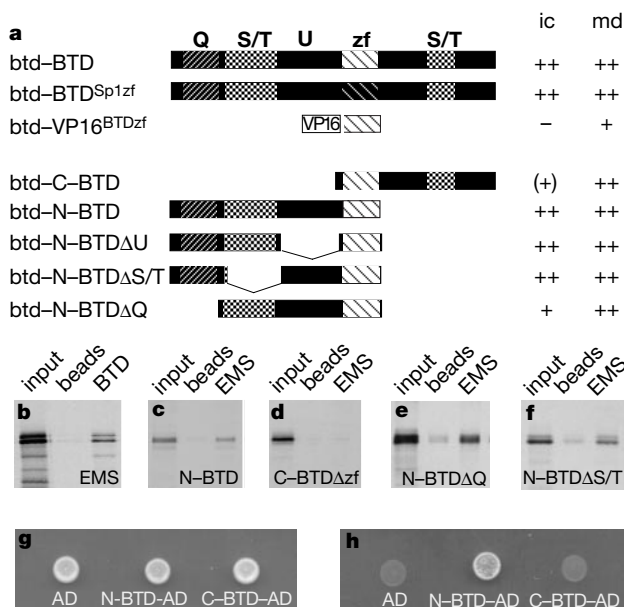


Figure 3 Transgene function in *btd* mutant embryos and interaction between EMS and BTd. **a**, Transgene constructs (left) and rescue of intercalary (ic) and mandibular (md) segments (criteria shown in Fig. 1) of transgene-bearing *btd*-mutant embryos. *btd*-BTD was reported previously^{7,13}. ++, complete rescue of head segments (see Fig. 1a–d) in all embryos ($n \geq 50$); +, partial rescue (subset of segment markers in high penetrance); (–), minimal rescue (a subset of segment markers in low penetrance); –, no rescue. Rescue was independent of the copy number of transgenes, indicating that the rescuing effects were not dose-dependent. *zf* (hatched boxes), zinc finger region shown to be essential for *btd* function⁷. **b–d**, BTd interacts with EMS *in vitro*. [³⁵S]methionine-labelled EMS interacts with Flag-tagged BTd (**b**). Radioactively labelled N terminus of BTd (N-BTD) interacts with EMS (**c**). The C terminus, which lacks the zinc finger domain (C-BTDΔ*zf*) does not bind EMS (**d**). **e, f**, N-terminal truncations N-BTDΔQ (**e**) and N-BTDΔS/T (**f**) interact with Flag-tagged EMS. Input refers to 10% of labelled proteins used for the binding reaction; beads refers to unreacted resin with labelled input protein. **g, h**, Yeast two-hybrid interactions between EMS and BTd. All yeast clones express the N terminus of EMS (lacking the homeodomain) fused to the GAL4-DNA-binding domain (EMSΔHD–BD). Indicated are the co-transformed GAL4-activation constructs: GAL4-activation domain (AD), N terminus of BTd (including zinc finger region) fused to AD (N-BTD–AD), C-terminal part of BTd (excluding the zinc finger region) fused to AD (C-BTDΔ*zf*–AD). **g**, Growth control on SD/– Leu/– Trp plates. **h**, Interaction assay on SD/– Leu/– Trp/– Ade plates using the stringent nutritional marker system based on an *ADE2* reporter (Clontech). Note that only N-BTD–AD interacts with EMSΔHD–BD.

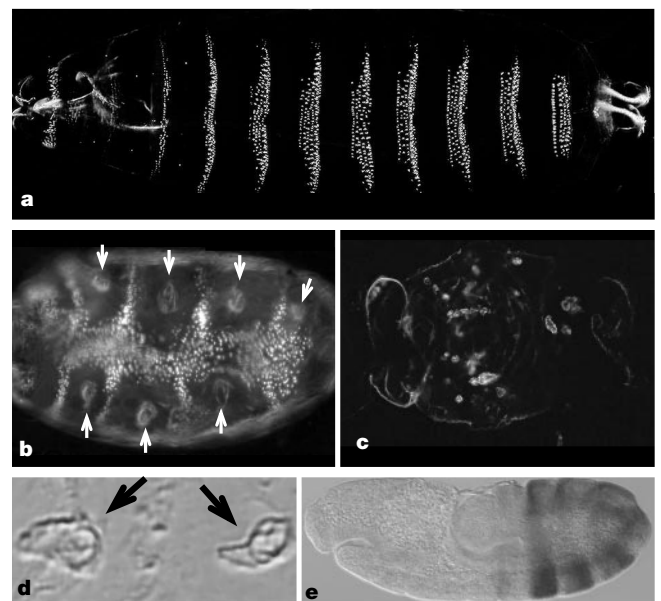


Figure 4 Ectopic *btd* expression lifts phenotypic suppression of *ems*. **a**, Cuticle pattern of a wild-type larva. **b–d**, Cuticle patterns of larvae that received heat shock-induced BTd at the time when EMS is expressed in a repeated pattern in the trunk region^{8,9}. Note a fusion of segments and the development of sclerotic head plates (arrows in **b**). About 5% of the embryos develop an extreme phenotype showing sclerotic plates (arrows in **d**) in the absence of segments (**c**; enlarged phase contrast micrograph in **d**). **e**, Ectopic BTd does not interfere with the expression of HOX-cluster genes as indicated by the normal pattern of *UBX* expression in the trunk region of stage 11 embryos. Note that the larvae shown in **b** and **c** are significantly smaller than wild-type larvae.

gap-like genes is therefore consistent with the proposed origin of the gene from the HOX-cluster³. By adopting BTΔ as a partner, EMS could escape phenotypic suppression by gnatho-cephalic HOX gene activities and specify the intercalary head segment identity. □

Methods

Drosophila strains

We used Oregon R, *btd*^{ΔG}, *svb*^{YPI7b} *btd*^{ΔG} (refs 7,19), homozygous lines of the transgenes described below and *hsp70-BTD/hsp70-BTD*; *ems*^{1/+} for heat-shock experiments in an *ems* mutant background. *svb btd* double mutant was used to identify *btd* mutant cuticles¹⁹.

Generation and analysis of transgenic animals

VP16^{BTDΔf}, N-BTD, C-BTD, N-BTDΔU, N-BTDΔS/T and N-BTDΔQ were constructed by polymerase chain reaction (PCR) and standard cloning procedures. N-BTD lacks amino acids 448–644 of the *btd* sequence⁷, C-BTD lacks 1–311, N-BTDΔU lacks 240–326 and 448–644, N-BTDΔS/T lacks 116–240 and 448–644, and N-BTDΔQ lacks 6–116 and 448–644. After sequencing, constructs were cloned into a P-element vector providing the 5.2 kilobases (kb) *btd* cis-acting element¹³. *btd*-EMS contains the 2.2 kb *Xba*I–*Eco*RV fragment of *ems* cDNA^{8,9}, UAS-EMS contains a 2.2 kb *Eco*RI fragment of *ems* cDNA in pUAST (ref. 15), and *hsp70-BTD* contains a 3.1 kb genomic *btd* *Bam*HI fragment in the *Bgl*II site of pCaSpeR-hs (ref. 20).

To generate transgenic flies, constructs were injected in *white* mutant embryos²¹. Except for N-BTDΔQ, at least two independent transgenic lines (balanced over CyO or TM3) were examined. Immunological stainings of embryos²² were performed with anti-β-galactosidase (Cappel), FP3.38 anti-UBX (ref. 23), 4D9 anti-EN (Developmental Studies Hybridoma Bank; University of Iowa)²² and 22C10 (ref. 24) primary antibodies using the Vectastain ABC Elite Kit (Vector). Homozygous mutant embryos were identified through blue balancers. Stained embryos were embedded (Canada Balsam, Sigma) or drawn into capillaries. Embryos (30-min collections) were heat-shocked (1 h; 37 °C) after 2 h of development (25 °C). Cuticle preparations¹⁹ and embryos were photographed with a Zeiss AxioPhot.

Protein binding assays

Full-length *ems* cDNA^{8,9} was cloned into a baculovirus transfer vector²⁵ to generate a flag-tag fusion construct for overproduction of EMS; the BTD and Sp1 constructs are described¹⁶. Recombinant baculovirus (Baculogold viral DNA, Pharmingen), expression and purification of Flag-epitope-tagged proteins from Sf9 cells were described²⁵. C-BTDΔf refers to the 880-bp carboxy-terminal *Bgl*II *Sp1 btd* fragment cloned into *Pvu*II-digested pRSETB (Invitrogen). For protein interaction studies, about 50 ng of Flag-epitope-tagged proteins (immobilized on Flag-M2 antibody resin; Eastman Kodak) were incubated (3 h, 4 °C) with [³⁵S]methionine-labelled proteins generated by the TNT-coupled *in vitro* transcription/translation system (Promega), washed extensively with 50 mM HEPES, pH 7.9, 25 mM MgCl₂, 40% glycerol, 0.8 M KCl, 1% Triton X-100, separated by SDS-PAGE and visualized by autoradiography.

Yeast two-hybrid assays were performed as described (Clontech manuals: Yeast Protocols Handbook; MATCHMAKER Two-Hybrid System 3). EMSΔHD–BD (residues 1–383) was generated by inserting an *Eco*RI/*Bam*HI fragment taken from EMSΔHD–AD into pGBK17. EMSΔHD–AD was PCR-amplified from *ems* cDNA^{8,9} (primers: EMS1F: 5'-CCCGAATTCATGACTAAGACGATTCGC-3'; EMS1R: 5'-CCGCCCCGGCTAGGGCACCAGGAACTTCC-3'), BTD1–AD (residues 1–217) and BTD2–AD (residues 105–424) were PCR-amplified from *btd* cDNA⁸ (primers: BTD1F: 5'-CGCGAATTCATGATCGATGCGGCCTGC-3'; BTD1R: 5'-GCCGGGCCCTACGCCGAGCTGCTGCTGCC-3' and BTD2F: 5'-GCCGAATTCCTATCCGGCTCGAGTTCC-3'; BTD2R: 5'-CGCGGGCCCTAGGCGGCCAGTACCTTCTTGC-3', respectively). PCR fragments were cloned into *Eco*RI/*Sma*I-digested pGADT7. N-BTD–AD (residues 1–424) was created by opening BTD1–AD with *Stu*I/*Bam*HI and inserting an *Stu*I/*Bam*HI fragment from BTD2–AD. C-BTD–AD (residues 405–645) was PCR-amplified (primers were BTD3F: 5'-GGCGGCATATGAGCGATCACCTCAGC-3'; BTD3R: 5'-CCCGGGCCCATCCTAGGCGGTGGC-3') and cloned into *Nde*I/*Sma*I-digested pGADT7.

Received 22 February; accepted 29 March 2000.

- Pankratz, M. & Jäckle, H. In *The Development of Drosophila melanogaster* (eds Bate, M. & Martinez Arias, A.) 467–516 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1993).
- Martinez Arias, A. In *The Development of Drosophila melanogaster* (eds Bate, M. & Martinez Arias, A.) 517–608 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1993).
- Manak, J. R. & Scott, M. P. A class act: conservation of homeodomain protein functions. *Dev. Suppl.* 61–77 (1994).
- Jürgens, G. & Hartenstein, V. In *The Development of Drosophila melanogaster* (eds Bate, M. & Martinez Arias, A.) 687–746 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1993).
- Cohen, S. M. & Jürgens, G. Mediation of *Drosophila* head development by gap-like segmentation genes. *Nature* 346, 482–485 (1990).
- Finkelstein, R., Smouse, D., Capaci, T. M., Spradling, A. C. & Perrimon, N. The *orthodenticle* gene encodes a novel homeo domain protein involved in the development of the *Drosophila* nervous system and ocular visual structures. *Genes Dev.* 4, 1516–1527 (1990).
- Wimmer, E. A., Jäckle, H., Pfeifle, C. & Cohen, S. M. A *Drosophila* homologue of human Sp1 is a head-specific segmentation gene. *Nature* 366, 690–694 (1993).
- Dalton, D., Chadwick, R. & McGinnis, W. Expression and embryonic function of *empty spiracles*: a *Drosophila* homeo box gene with two patterning functions on the anterior-posterior axis of the embryo. *Genes Dev.* 3, 1940–1956 (1989).

- Walldorf, U. & Gehring, W. J. *Empty spiracles*, a gap gene containing a homeobox involved in *Drosophila* head development. *EMBO J.* 11, 2247–2259 (1992).
- Macías, A. & Morata, G. Functional hierarchy and phenotypic suppression among *Drosophila* homeotic genes: the *labial* and *empty spiracles* genes. *EMBO J.* 15, 334–343 (1996).
- Lewis, E. B. A gene complex controlling segmentation in *Drosophila*. *Nature* 276, 565–570 (1978).
- Duboule, D. & Morata, G. Colinearity and functional hierarchy among genes of the homeotic complexes. *Trends Genet.* 10, 358–364 (1994).
- Wimmer, E. A., Cohen, S. M., Jäckle, H. & Desplan, C. *buttonhead* does not contribute to a combinatorial code proposed for *Drosophila* head development. *Development* 124, 1509–1517 (1997).
- Gallitano-Mendel, A. & Finkelstein, R. Ectopic *orthodenticle* expression alters segment polarity gene expression but not head segment identity in the *Drosophila* embryo. *Dev. Biol.* 199, 125–137 (1998).
- Janody, F., Reischl, J. & Dostatni, N. Persistence of Hunchback in the terminal region of the *Drosophila* blastoderm embryo impairs anterior development. *Development* 127, 1573–1582 (2000).
- Schöck, F., Sauer, F., Jäckle, H. & Purnell, B. A. *Drosophila* head segmentation factor *Buttonhead* interacts with the same TATA box-binding protein-associated factors and *in vivo* DNA targets as human Sp1 but executes a different biological program. *Proc. Natl Acad. Sci. USA* 96, 5061–5065 (1999).
- Sadowski, I., Ma, J., Triezenberg, S. & Ptashne, M. GAL4-VP16 is an unusually potent transcriptional activator. *Nature* 335, 563–564 (1988).
- Jones, B. & McGinnis, W. The regulation of *empty spiracles* by *Abdominal-B* mediates an abdominal segment identity function. *Genes Dev.* 7, 229–240 (1993).
- Wimmer, E. A., Frommer, G., Purnell, B. A. & Jäckle, H. *buttonhead* and *D-Sp1*: a novel *Drosophila* gene pair. *Mech. Dev.* 59, 53–62 (1996).
- Thummler, C. S. & Pirota, V. New pCaSpeR P-element vectors. *Drosoph. Inf. Serv.* 71, 150 (1992).
- Rubin, G. M. & Spradling, A. C. Genetic transformation of *Drosophila* with transposable element vectors. *Science* 218, 348–353 (1982).
- Patel, N. H. *et al.* Expression of *engrailed* proteins in arthropods, annelids, and chordates. *Cell* 58, 955–968 (1989).
- White, R. A. H. & Wilcox, M. Protein products of the Bithorax complex in *Drosophila*. *Cell* 39, 163–171 (1984).
- Zipursky, S. L., Venkatesh, T. R., Teplow, D. B. & Benzer, S. Neuronal development in the *Drosophila* retina: monoclonal antibodies as molecular probes. *Cell* 36, 15–26 (1984).
- Sauer, F., Hansen, S. K. & Tjian, R. Multiple TAFs directing synergistic activation of transcription. *Science* 270, 1783–1788 (1995).
- Schmidt-Ott, U., González-Gaitán, M., Jäckle, H. & Technau, G. M. Number, identity, and sequence of the *Drosophila* head segments as revealed by neural elements and their deletion patterns in mutants. *Proc. Natl Acad. Sci. USA* 91, 8363–8367 (1994).
- Rogers, B. T. & Kaufman, T. C. Structure of the insect head as revealed by the EN protein pattern in developing embryos. *Development* 122, 3419–3432 (1996).

Acknowledgements

We thank G. Dowe for sequencing; M. González-Gaitán, G. Vorbrüggen and R. Rivera-Pomar for discussions; C. Klämbt for the 22C10 antibody; and F. Janody and N. Dostatni for the maternal Gal4 driver. The work was supported by the Human Frontier Science Organization (H.J.) and by fellowships of the Fonds der Chemischen Industrie (F.S.), the Alexander-von-Humboldt Stiftung (B.A.P.) and the Boehringer Ingelheim Fonds (J.R.).

Correspondence and requests for materials should be addressed to H.J. (e-mail: hjaeckl@gwdg.de).

Blockade of RAGE–amphoterin signalling suppresses tumour growth and metastases

Akihiko Taguchi*, David C. Blood*, Gustavo del Toro*, Anthony Canet*, Daniel C. Lee*, Wu Qu*, Nozomu Tanji*, Yan Lu*, Evanthal Lalla*, Caifeng Fu*, Marion A. Hofmann*, Thomas Kislinger*, Mark Ingram*, Amy Lu*, Hidekazu Tanaka†, Osamu Hori‡, Satoshi Ogawa ‡, David M. Stern* & Ann Marie Schmidt*

* College of Physicians & Surgeons, Columbia University, New York, New York 10032, USA

† Osaka University School of Medicine, Osaka 565-0871, Japan

‡ Kanazawa University School of Medicine, Kanazawa 920-8640, Japan

The receptor for advanced glycation end products (RAGE), a multi-ligand member of the immunoglobulin superfamily of cell surface molecules^{1–2}, interacts with distinct molecules implicated in homeostasis, development and inflammation, and certain diseases such as diabetes and Alzheimer's disease^{3–8}. Engagement of RAGE by a ligand triggers activation of key cell signalling

pathways, such as p21^{ras}, MAP kinases, NF- κ B and cdc42/rac, thereby reprogramming cellular properties^{9–11}. RAGE is a central cell surface receptor for amphoterin, a polypeptide linked to outgrowth of cultured cortical neurons derived from developing brain^{3,12–15}. Indeed, the co-localization of RAGE and amphoterin at the leading edge of advancing neurites indicated their potential contribution to cellular migration, and in pathologies such as tumour invasion. Here we demonstrate that blockade of RAGE–amphoterin decreased growth and metastases of both implanted tumours and tumours developing spontaneously in susceptible mice. Inhibition of the RAGE–amphoterin interaction suppressed activation of p44/p42, p38 and SAP/JNK MAP kinases; molecular effector mechanisms importantly linked to tumour proliferation, invasion and expression of matrix metalloproteinases^{16–23}.

Amphoterin and RAGE are expressed in a range of cell types^{15,24}, and RAGE–amphoterin modulates cellular invasive properties in developing neurons. Thus, it was logical to examine the potential role of these molecules in tumour biology. Rat C6 glioma cells¹⁵ provided an ideal starting point for these studies, as immunoblotting of C6 glioma cell lysates demonstrated expression of both RAGE (Fig. 1a, lane 3) and amphoterin (Fig. 1b, lane 3). Confocal microscopy confirmed the co-localization of RAGE and amphoterin in these cells (see Fig. 1 in Supplementary Information). The absence of another RAGE ligand, EN-RAGE⁸, in C6 glioma (Fig. 1c, lanes 3 and 5) led us to pursue potential roles for amphoterin and RAGE in tumour behaviour.

We focused on strategies to inhibit the function of RAGE and amphoterin *in vivo*. The extracellular region of RAGE is composed of one 'V'-type followed by two 'C'-type immunoglobulin-like domains and comprises the soluble, ligand-binding domain². Following the extracellular and transmembrane spanning domains is a short, highly charged cytosolic tail, which is essential for RAGE-dependent intracellular signalling and subsequent cellular activation⁸. Therefore, four strategies for blocking RAGE–amphoterin-mediated cellular stimulation *in vivo* are: (1) administration of soluble, extracellular ligand-binding domain of RAGE, sRAGE; (2) administration of blocking F(ab')₂ fragments derived from anti-RAGE and/or anti-amphoterin IgG; (3) generation of stably transfected C6 glioma expressing a RAGE mutant devoid of the cytosolic tail, so that the truncated form of the receptor is present on the cell surface and competent for ligand binding, but exerts a dominant-negative effect on RAGE signalling; and (4) generation of stably transfected C6 glioma expressing sRAGE.

Administration of sRAGE once daily^{4,5} to immunocompromised (athymic nude) mice upon injection of rat C6 glioma cells caused dose-dependent decreases in tumour volume (Fig. 2a). As amphoterin and RAGE were both present in the tumour bed, we assessed their effects by using monospecific polyclonal rabbit F(ab')₂ fragments prepared from antibodies to RAGE and/or amphoterin. These antibody fragments were administered to immunocompromised (severe combined immunodeficiency; SCID) mice from the time of inoculation with C6 glioma cells. Compared with SCID mice receiving nonimmune F(ab')₂ or anti-EN-RAGE F(ab')₂, animals treated with anti-amphoterin or anti-RAGE F(ab')₂ had a significant reduction in tumour volume after 21 days (Fig. 2b). Simultaneous treatment with both anti-RAGE and anti-amphoterin F(ab')₂ resulted in greater reduction in tumour volume compared with treatment with the control F(ab')₂ or either antibody alone (Fig. 2b).

A limitation of these modalities was the possibility that such proteins administered intraperitoneally would not reach sufficiently high levels in the tumour bed to inhibit the receptor. Therefore, we stably transfected rat C6 glioma cells with tail-deletion RAGE (T), soluble RAGE (S) and full-length RAGE (F), or mock-transfected them (M). Three independent clones from each transfection were implanted into immunocompromised (athymic nude) mice to evaluate the role of RAGE in tumour growth. Analysis of the resulting tumours by immunoblotting revealed the following

increase in RAGE expression compared with mock-transfected clones: clones F1–3: 11.7-, 9.0- and 22.6-fold; clones T1–3: 20.2-, 7.7- and 9.4-fold; and clones S1–3: 3.4-, 7.0- and 8.4-fold. Compared with mock-transfectants (*n* = 17) (Fig. 2c, d), about a fivefold increase in tumour volume was observed in cells overexpressing full-length RAGE on day 21 (*n* = 18) (Fig. 2c, e). Tumour volume was decreased about threefold in tumours derived from C6 glioma clones transfected with the tail-deletion mutant (*n* = 21) (Fig. 2c, f). Tumour volume also decreased (about 6.5-fold) in neoplasms derived from sRAGE-transfected C6 glioma (*n* = 18) (Fig. 2c, g). These results indicate the involvement of RAGE-mediated cellular activation in tumour growth and phenotype.

To delineate the mechanisms by which blockade of RAGE suppressed tumour growth, we analysed implanted tumours over time. On days 1, 3 and 7, mock- and full-length RAGE-transfected C6 glioma had begun to grow and invade surrounding muscle and connective tissue. In contrast, transfectants expressing tail-deletion RAGE or sRAGE on days 1, 3 and 7 were limited to the immediate area where they had been injected. Only at day 14 did tail-deletion RAGE transfectants and sRAGE-transfectants begin to invade the surrounding tissues (see Fig. 2 in Supplementary Information). Consistent with these observations, cells expressing glial fibrillary acidic protein (GFAP) at the centre of the tumour increased in area in full-length RAGE-transfected clones on days 3, 7, and 14 as compared with mock-transfected tumours (Fig. 3a). In contrast, tumour cells expressing tail-deletion RAGE or sRAGE had a markedly decreased area on days 1, 3, 7 and 14 after implantation (Fig. 3a).

Diminished tumour size in the setting of RAGE blockade indicated that RAGE might modulate cellular proliferation and/or cell death. Compared with mock-transfectants, C6 glioma expressing full-length RAGE exhibited enhanced incorporation of 5-bromo-2'-deoxyuridine (BrdU) on days 1 and 3 after implantation (Fig. 3b). In contrast, tail-deletion RAGE and sRAGE transfectants exhibited a decrease in incorporation of BrdU on days 1, 3 and 7 compared with mock-transfected clones (Fig. 3b). However, rates of apoptosis were low (0.2–0.5% from days 1–14) and did not change among the various transfectants. Similar results were obtained when sRAGE was administered intraperitoneally; on days 1, 3 and 7 after injection, decreased proliferation was noted in the presence of sRAGE (Fig. 3c) and the apoptotic rate was not different from that observed in mice receiving murine serum albumin (MSA) (data not shown). Consistent with the versatility of tumour cells was their ability to escape the suppression of cell growth associated with blockade of RAGE subsequent to day 14. This is when most tail-deletion and sRAGE transfectants, and tumour cells subjected to parenteral administration of sRAGE were first measurable. Subsequently, tumours from all groups escaped latency induced by blockade of RAGE and displayed equivalent growth rates (data not shown).

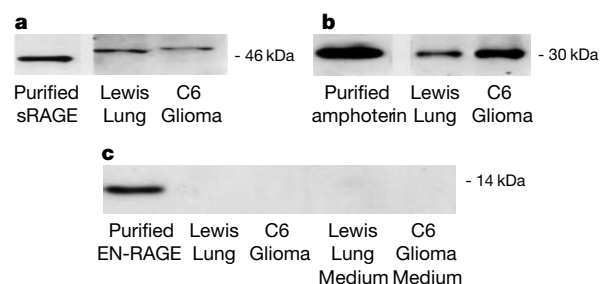


Figure 1 RAGE and amphoterin are expressed in tumour cells. **a–c**, Cell lysates or supernatant (medium) were prepared from cultured tumour cells and subjected to immunoblotting for RAGE (**a**), amphoterin (**b**), or EN-RAGE (**c**).

Amphotericin may provide a surface for assembly of fibrinolytic complexes leading to generation of plasmin, a central molecule in activation of matrix metalloproteinases (MMP)^{14,15}, indicating a mechanism by which RAGE–amphotericin might enhance invasive properties of tumour cells. On day 21 after implantation, those tumours overexpressing full-length RAGE had increased activity of MMP-9 and MMP-2 compared with mock-transfected C6 glioma (Fig. 3d). Tumours raised from tail-deletion RAGE or sRAGE in transfected clones, however, showed diminished MMP-9 and MMP-2 activity compared with those from mock-transfected clones (Fig. 3d).

Consistent with these *in vivo* observations, tail-deletion RAGE and sRAGE transfectants plated on amphotericin-coated matrices had diminished proliferation compared with mock-transfected C6 glioma (Fig. 4a). Further, overexpression of full-length RAGE in C6 glioma grown on amphotericin enhanced proliferation (Fig. 4a). These findings were selective for growth on amphotericin, as pro-

liferation did not vary between mock-, full-length RAGE-, tail-deletion RAGE- and sRAGE-transfected C6 glioma grown on substrates with adsorbed bovine serum albumin (data not shown). Similar inhibition of cellular proliferation was observed when sRAGE was added to wild-type C6 glioma on amphotericin-coated matrices (data not shown).

As our findings indicated that RAGE and amphotericin mediated invasion and migration of implanted C6 glioma, we assessed these properties *in vitro*. Compared with mock-transfected C6 glioma, those cells with full-length RAGE demonstrated enhanced invasion through Matrigel (Fig. 4b). In contrast, transfected clones bearing the tail-deletion RAGE mutant or overexpressing sRAGE had diminished invasion (Fig. 4b). In the presence of sRAGE, anti-RAGE F(ab')₂ or anti-amphotericin F(ab')₂, significant suppression of invasion was observed (Fig. 4b). Similarly, blockade of RAGE–amphotericin suppressed migration of C6 glioma *in vitro* (data not shown). Also, mock- and full-length RAGE transfected C6 glioma

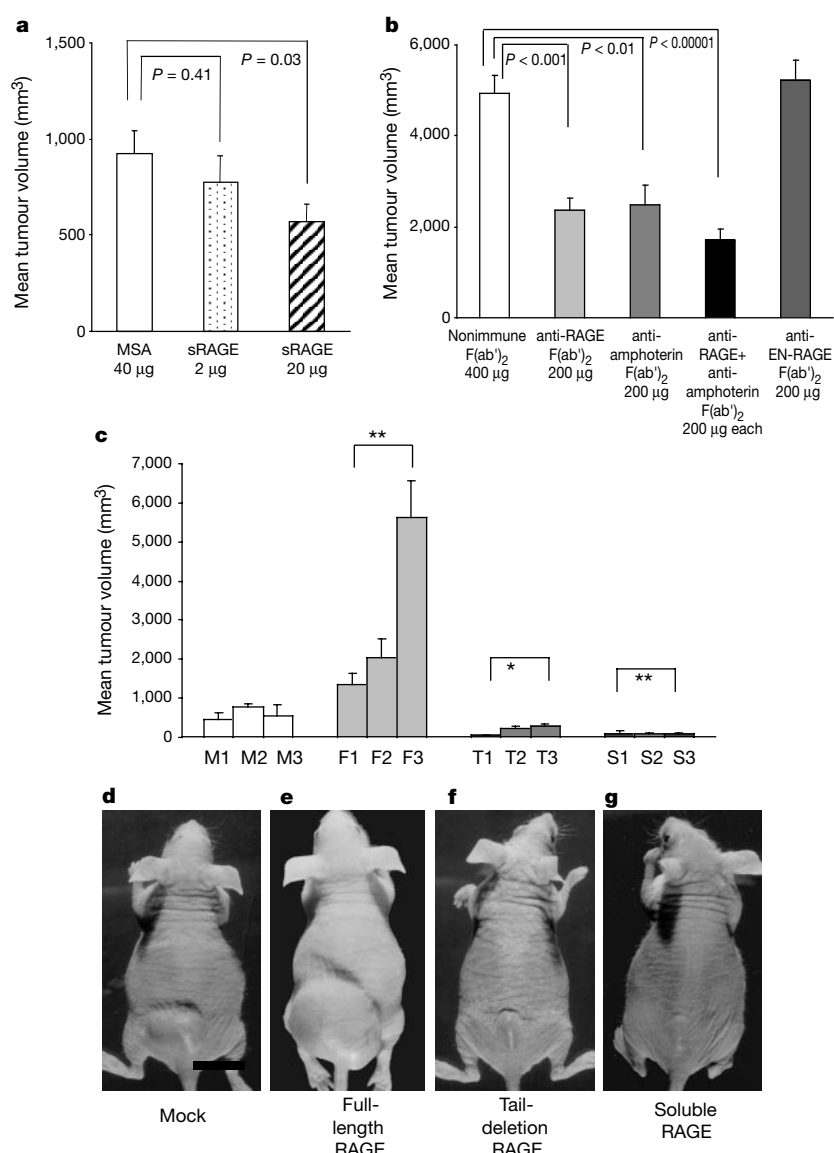


Figure 2 RAGE–amphotericin blockade suppresses growth of implanted C6 glioma. **a**, C6 glioma cells were implanted into immunocompromised mice and mean tumour volume on day 21 is shown, $n = 10$ per group. MSA, murine serum albumin; sRAGE, soluble RAGE. **b**, Mice with severe combined immunodeficiency (SCID) were injected with C6 glioma cells and the indicated F(ab')₂ fragments. Mean tumour volume on day 21 is shown; $n = 7$

per group. **c**, Mock or RAGE/RAGE mutant C6 glioma cells were injected into immunocompromised mice and mean tumour volume on day 21 is shown. Asterisk, $P < 0.01$, double asterisk $P < 0.001$ versus mock-transfected clones. **d–g**, Representative photographs on day 21 of mice bearing the indicated transfected clone are shown. Scale bar, 1 cm.

spread readily on amphotericin-coated matrix (Fig. 4c, d), tail-deletion RAGE and sRAGE-transfected cells (Fig. 4e, f) had a marked reduction in their ability to extend processes necessary to migrate into the surrounding matrix.

To examine the molecular mechanisms underlying RAGE–amphotericin-mediated effects on tumour properties, we focused on the MAP kinase family of signalling effector molecules, as these mediators are involved in cellular proliferation, invasion and activation of MMPs^{16–23}. Although activation of p44/p42, p38 and SAP/JNK MAP kinases was enhanced in full-length RAGE transfectants plated on amphotericin, a marked reduction in activation of these kinases on amphotericin was noted in tail-deletion RAGE and sRAGE transfectants (Fig. 4g–i, left). In contrast, activation of MAP kinases did not differ among the transfectants grown on BSA, suggesting the specificity of amphotericin–RAGE interaction in activating these key cell signalling molecules (Fig. 4g–i, right).

Another potent mechanism by which growth and spread of tumour cells may be modulated is by suppression of neovascularization. To explore whether RAGE blockade impaired this process, basic fibroblast growth factor (b-FGF)-laden pellets were placed into a corneal pocket, and new capillary growth from the corneal limbus was observed²⁵. New vessel growth was not different in C57BL/6J mice treated with either MSA or sRAGE for five days ($3.1 \pm 0.3 \text{ mm}^2$ and $2.7 \pm 0.2 \text{ mm}^2$ angiogenic area, respectively; $P > 0.05$).

As the ability of tumour cells to proliferate and invade beyond homeostatic boundaries is a central means by which the host succumbs to tumour, it was important to ascertain whether RAGE might contribute to metastases. We used the Lewis lung carcinoma model, in which distant metastases flourish upon removal of the primary tumour. Both RAGE and amphotericin were present in Lewis lung carcinoma cells (Fig. 1a, b), and could be localized to the cell surface by immunocytochemistry (not shown). We first prepared stably transfected Lewis lung carcinoma

cells overexpressing sRAGE. However, following implantation into C57BL/6J mice, marked suppression of local tumour growth resulted (data not shown), thereby abrogating the usefulness of this model. As an alternative strategy, sRAGE was administered just before and after resection of primary tumours resulting from inoculation with wild-type Lewis lung carcinoma cells. Compared with MSA, animals receiving sRAGE at $100 \mu\text{g}$ per day demonstrated a marked decrease in the number of lung surface metastases (8.7 ± 1.4 and 1.0 ± 0.3 , respectively; $P < 0.0001$) and metastatic burden as judged by lung weight ($385.6 \pm 39.8 \text{ mg}$ and $188.7 \pm 6.7 \text{ mg}$, respectively; $P < 0.001$). Lung surface metastases observed in MSA-treated mice were virtually undetectable in mice treated with sRAGE (see Fig. 3 in Supplementary Information).

To study early RAGE-mediated events in metastasis, fluorescently labelled Lewis lung carcinoma cells were intravenously injected into mice; 24 h later, cell number in the lungs was assessed. Compared with MSA-treated mice, a significant reduction in tumour cell number under high-powered field was noted in the presence of sRAGE (1.1 ± 0.03 and 0.4 ± 0.02 , respectively; $P < 0.0001$). On day 14, the number of lung surface metastases was reduced in mice treated with sRAGE compared with those receiving MSA (13.0 ± 1.5 and 47.6 ± 4 , respectively; $P < 0.00001$), as was lung weight ($194.8 \pm 8.8 \text{ mg}$ and $405.8 \pm 25.3 \text{ mg}$, respectively; $P < 0.00001$).

The critical test of our hypotheses was whether blockade of RAGE–amphotericin might affect endogenous growth of tumours. We tested these concepts in spontaneously appearing papillomas in mice overexpressing v-Ha-ras transgene. After stimulation of the skin with a tumour promoter such as phorbol 12-myristate 13-acetate (PMA), papillomas appear in >90% of mice within six weeks²⁶.

RAGE and amphotericin were highly expressed in endogenously formed papillomas in PMA-treated transgenic mice (Fig. 5b, c). To

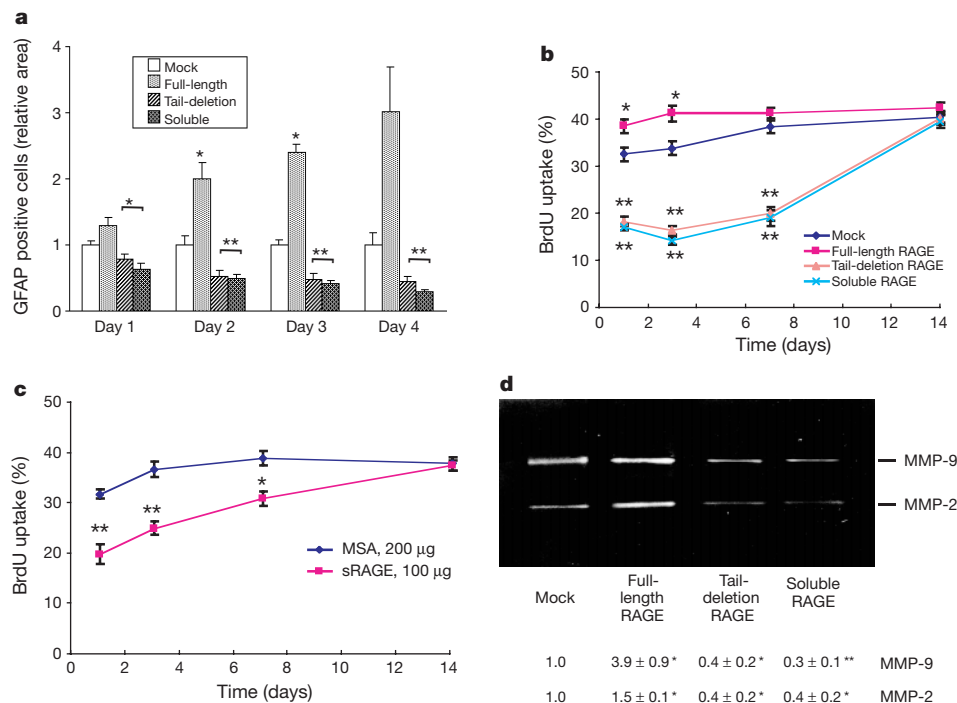


Figure 3 Blockade of RAGE suppresses tumour proliferation and expression of MMPs. **a**, Mean area at the centre of tumours raised from the indicated clones of C6 glioma was determined by immunohistochemical determination of GFAP-expressing tumour cells; $n = 5$ mice. **b,c**, The indicated C6 glioma transfectants (**b**) or C6 glioma in the presence of MSA or sRAGE (**c**) were implanted into mice; 1 h before sacrifice, mice were injected with BrdU; tissue was retrieved and subjected to immunohistochemistry to assess

incorporation of BrdU. **d**, Tumours ($n = 3$ per clone) were retrieved 21 days after implantation and zymographic determination of MMP9 or MMP2 activity performed. Results of densitometric analysis are indicated; 1.0 is arbitrarily assigned to density of bands in mock-transfected tumours. In **a–d**, asterisk, $P < 0.05$, double asterisk, $P < 0.01$ versus mock or MSA.

study RAGE–amphoterin blockade, sRAGE or MSA was administered to transgenic mice. After six weeks, marked suppression of papilloma formation was noted in sRAGE-treated transgenic animals compared with treatment with MSA (Fig. 5e, f; mean papillomas per animal, 3.8 ± 1.0 and 0.3 ± 0.3 , respectively; $P < 0.01$).

Consistent with earlier observations, incorporation of BrdU was attenuated in papillomas studied from sRAGE-treated transgenic

mice compared with those retrieved from mice receiving MSA ($22 \pm 2\%$ and $39 \pm 2\%$, respectively; $P < 0.0001$), and no differences in apoptotic rates were observed between papillomas in sRAGE-treated and MSA-treated mice ($0.1 \pm 0.1\%$ and $0.2 \pm 0.1\%$, respectively; $P > 0.05$).

In embryonic development, amphoterin and RAGE co-localize at the leading edge of advancing neurites, indicating a role in neuronal

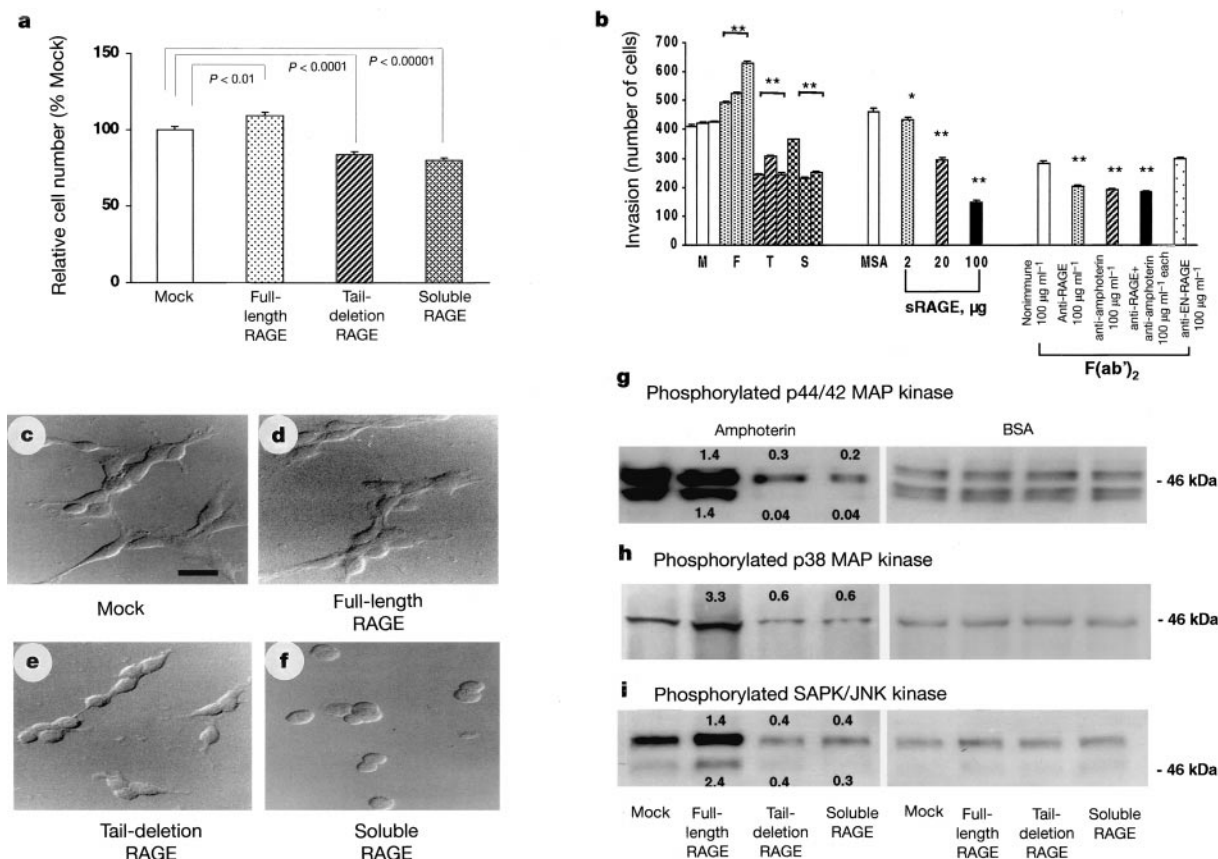


Figure 4 Effects of RAGE–amphoterin blockade on C6 glioma: *in vitro* analyses. **a**, The indicated clones of C6 glioma were grown on matrix coated with amphoterin. Cell number was quantified 72 h later and results reported as relative cell number, compared with that observed in mock-transfectants. **b**, The number of cells invading Matrigel is shown. Asterisk, $P < 0.05$; double asterisk $P < 0.00001$ versus respective control. **c–f**, Appearance of the indicated mock- or RAGE/RAGE C6 glioma mutants was assessed

on amphoterin. Scale bar, 20 μm . **g–i**, The indicated clones of C6 glioma were plated on amphoterin (left) or BSA (right) for 90 mins; cells were retrieved and immunoblotting performed to assess activation of p44/p42 (**g**), p38 (**h**) and SAP/JNK MAP kinases (**i**). Densitometric analysis for kinase activation on amphoterin compared with mock-transfected clones (1.0) is indicated.

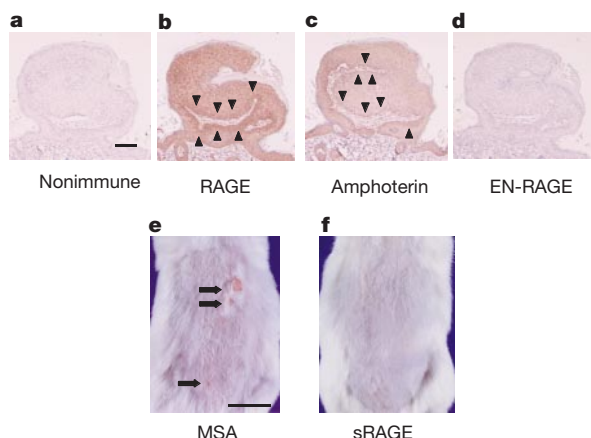


Figure 5 Blockade of RAGE suppresses endogenous growth of papillomas. **a–d**, After six weeks of local application of PMA, papillomas forming in transgenic mice overexpressing v-Ha-ras were examined by immunohistochemistry using the indicated IgG. Arrowheads indicate the sites of most intense immunostaining for RAGE and amphoterin; these areas

co-localized with sites of highest levels of BrdU incorporation (not shown). Scale bar, 100 μm . **e, f**, After six weeks, increased numbers of papillomas (arrows) were observed in mice receiving MSA (**e**; $n = 8$) versus sRAGE (**f**; $n = 7$). Scale bar, 1 cm.

development^{3,15}. Here we extend these concepts to tumour growth and metastasis, and demonstrate RAGE–amphoterin involvement in cellular proliferation and invasiveness and that it is also central for tumour growth and spread. Blockade of RAGE–amphoterin in the tumour bed forces cells into a period of dormancy²⁷, characterized by diminished proliferation, invasion and MMP activity; properties linked to the MAP kinase family of signal transduction effector molecules. We have developed a new model in tumour biology and identified RAGE as a target for therapeutic strategies to suppress local tumour growth and distant metastases. These new therapies will probably enhance the benefit of other anti-tumour strategies, such as those designed to diminish neovascularization, proliferation and evasion of the host immune response. □

Methods

Immunoblots

Rat C6 glioma and murine Lewis lung carcinoma cell lines were purchased from the American Type Culture Corporation (ATCC). Immunoblotting was performed with 30 µg cell extract⁸ or cellular supernatant and using rabbit anti-RAGE IgG³, rabbit anti-rat amphoterin IgG³, rabbit anti-EN-RAGE IgG⁸ or equal amounts of nonimmune IgG (25 µg ml⁻¹). In all cases, control recombinant proteins (0.5 µg) were murine soluble RAGE, rat amphoterin and bovine EN-RAGE. All control proteins crossreact with murine and rat antigens.

Confocal microscopy

C6 glioma were fixed with paraformaldehyde (2%) and confocal microscopy (Zeiss) was performed using anti-RAGE and anti-amphoterin IgG. Secondary antibodies included FITC-conjugate (for amphoterin) and TRITC-conjugate (for RAGE) (Sigma; 1:200 dilution in both cases).

Immunohistochemistry

Implanted C6 glioma were excised and fixed with formalin; paraffin-embedded sections (5 µm thick) were prepared from the exact centre of the tumours, and subjected to immunohistochemistry with anti-glial fibrillary acidic protein Ig (Sigma). Microscopic images (Zeiss) of GFAP-stained sections were scanned into a computer and image analysis for determination of area using software provided by MediaCybernetics⁵. For endogenously forming papillomas, sections were prepared as above and subjected to immunohistochemistry using rabbit nonimmune IgG, anti-RAGE IgG, anti-amphoterin IgG or anti-EN-RAGE IgG (2 µg ml⁻¹ each).

Stably transfected C6 glioma complementary DNAs for human full-length, cytosolic tail deletion and soluble RAGE² were inserted into the pcDNA3 vector (Invitrogen). C6 rat glioma cells were transfected by using Lipofectamine (Life Technologies). Cells were selected in the presence of geneticin (G418), 1.5 mg ml⁻¹ (Life Technologies), and individual clones were isolated by limiting dilution. Mock-transfectants contained vector alone.

Local tumour growth

Rat C6 glioma cells (1 × 10⁵ in 0.1 ml PBS) were injected into the dorsal midline of female NCR immunocompromised mice, aged 4–6 weeks (Taconic Farms, Germantown, NY). Alternatively, rat C6 glioma cells were injected into the dorsal midline of female mice with severe combined immunodeficiency (SCID; Taconic Farms). Tumours were measured with calipers and the volume was calculated: $V = \pi \times h(h^2 + 3a^2)/6$, where h = height of the tumour segment; a = (length + width of the tumour)/4; and V = volume of the tumour. In all cases, viability of injected cells was > 95% by exclusion of trypan blue upon injection into mice. In certain mice, one hour before sacrifice, 1 mg BrdU (Sigma) was injected by intraperitoneal administration. Tumour tissue was retrieved, fixed in formalin (10%) and paraffin-embedded sections were prepared. Sections were stained with haematoxylin and eosin, or were simultaneously immunostained for detection of BrdU (anti-BrdU antibody; Accurate, Westbury, NY) and for *in situ* apoptosis detection (Trevigen, Gaithersburg, MD).

Tumour metastases

Lewis lung murine carcinoma cells (2 × 10⁶ in 0.1 ml PBS) were injected into the dorsal midline of male, 6–8-week-old C57BL/6 mice (Jackson Laboratories, Bar Harbor, ME). Primary tumours were surgically excised when tumour volume was 1,500 mm³ (day 14). For three days before sacrifice, mice received sRAGE or MSA once daily, 21 days after removal of primary tumour. Weight of the lungs and numbers of lung surface metastases were determined under ×4 magnification using an Olympus microscope after intratracheal injection of India Ink (15%). In other experiments, Lewis lung carcinoma cells were labelled with Vybrant CFDA (Molecular Probes, Eugene, OR). 2 × 10⁵ cells in 0.1 ml PBS were injected intravenously into C57BL/6 mice and animals were sacrificed 24 h or 14 days after injection. Lungs were removed, fixed and paraffin-embedded sections prepared (5 µm thick). Fluorescence microscopy was used to identify tumour cells; 60 sections were assessed per mouse.

Endogenous tumour formation

Male transgenic mice (hemizygous) carrying the activated v-Ha-ras transgene²⁶ were purchased from Taconic Farms. At age 8 weeks, mice were housed in single cages and their backs shaved (4 × 2 cm). PMA (Sigma), 5 µg dissolved in acetone (200 µl), was administered locally to the shaved area twice weekly for 6 weeks.

Tumour properties

To assess proliferation, the indicated C6 glioma cells (1 × 10³ cells per well) were incubated on culture wells coated with amphoterin (10 µg ml⁻¹) or BSA (20 µg ml⁻¹). MSA or sRAGE was added, and three days later cell number was assessed using the CyQUANT Cell Proliferation Assay kit (Molecular Probes). Absence of cell death was documented by exclusion of trypan blue. Invasion and migration assays were performed as described^{28,29}.

For the assessment of MMP-2 and 9 activity³⁰, tumour tissue was retrieved on day 21. Equal amounts of protein were subjected to electrophoresis on gelatin-laden gels (0.1%) (Novex) and results normalized to the weight of the tumour.

To examine the properties of the tumours grown on amphoterin, C6 glioma cells, 3 × 10⁴ cells per well, were added to plastic dishes (Nunc, Naperville, IL) coated with purified amphoterin (10 µg ml⁻¹) or BSA (20 µg ml⁻¹). After 18 h, cells were photographed under phase contrast microscopy. In other studies, cells were retrieved 90 min after plating, and extracts (30 µg protein) subjected to electrophoresis and immunoblotting to detect phospho-p44/42 MAP kinase (Thr202/Tyr204), phospho-p38 MAP kinase (Tyr180/Tyr182), or phospho-SAPK/JNK (Thr183/Tyr185) (New England BioLabs, Beverly, MA).

Data analysis

In all experiments, mean ± standard error is reported. Statistical comparisons among groups were determined using one-way analysis of variance (ANOVA); where indicated, individual comparisons were performed using Students' *t*-test.

Received 16 December 1999; accepted 16 March 2000.

- Schmidt, A. M. *et al.* Isolation and characterization of binding proteins for advanced glycosylation endproducts from lung tissue which are present on the endothelial cell surface. *J. Biol. Chem.* **267**, 14987–14997 (1992).
- Neeper, M. *et al.* Cloning and expression of RAGE: a cell surface receptor for advanced glycosylation end products of proteins. *J. Biol. Chem.* **267**, 14998–15004 (1992).
- Hori, O. *et al.* The receptor for advanced glycation endproducts (RAGE) is a cellular binding site for amphoterin: mediation of neurite outgrowth and coexpression of RAGE and amphoterin in the developing nervous system. *J. Biol. Chem.* **270**, 25752–25761 (1995).
- Wautier, J.-L. *et al.* Receptor-mediated endothelial cell dysfunction in diabetic vasculopathy: soluble receptor for advanced glycation endproducts blocks hyperpermeability. *J. Clin. Invest.* **97**, 238–243 (1996).
- Park, L. *et al.* Suppression of accelerated diabetic atherosclerosis by soluble receptor for AGE (sRAGE). *Nature Med.* **4**, 1025–1031 (1998).
- Yan, S. D. *et al.* RAGE and amyloid beta peptide neurotoxicity in Alzheimer's disease. *Nature* **382**, 685–691 (1996).
- Yan, S. D. *et al.* Amyloid-beta peptide-RAGE interaction elicits neuronal expression of M-CSF: a proinflammatory pathway in Alzheimer's disease. *Proc. Natl Acad. Sci. USA* **94**, 5296–5301 (1997).
- Hofmann, M. A. *et al.* RAGE mediates a novel proinflammatory axis: a central cell surface receptor for S100/calgranulin polypeptides. *Cell* **97**, 889–901 (1999).
- Yan, S.-D. *et al.* Enhanced cellular oxidant stress by the interaction of advanced glycation endproducts with their receptors/binding proteins. *J. Biol. Chem.* **269**, 9889–9897 (1994).
- Lander, H. L., Tauras, J. M., Ogiste, J. S., Moss, R. A. & Schmidt, A. M. Activation of the receptor for advanced glycation endproducts triggers a MAP kinase pathway regulated by oxidant stress. *J. Biol. Chem.* **272**, 17810–17814 (1997).
- Huttunen, H. J., Fages, C. & Rauvala, H. Receptor for advanced glycation endproducts (RAGE)-mediated neurite outgrowth and activation of NF-κB require the cytoplasmic domain of the receptor but different downstream signaling pathways. *J. Biol. Chem.* **274**, 19919–19924 (1999).
- Rauvala, H. & Pihlaskari, R. Isolation and some characteristics of an adhesive factor of brain that enhances neurite outgrowth in central neurons. *J. Biol. Chem.* **262**, 16625–16635 (1987).
- Rauvala, H. *et al.* The adhesive and neurite-promoting molecule p30: analysis of the amino-terminal sequence and production of antipeptide antibodies that detect p30 at the surface of neuroblastoma cells and of brain neurons. *J. Cell. Biol.* **107**, 2293–22305 (1988).
- Parkkinen, J. & Rauvala, H. Interactions of plasminogen and tissue plasminogen activator (t-PA) with amphoterin. *J. Biol. Chem.* **266**, 16730–16735 (1991).
- Parkkinen, J. *et al.* Amphoterin, the 30-kDa protein in a family of HMGI-type polypeptides. Enhanced expression in transformed cells, leading edge localization, and interactions with plasminogen activation. *J. Biol. Chem.* **268**, 19726–19738 (1993).
- Brunet, A. *et al.* Nuclear translocation of p42/p44 mitogen-activated protein kinase is required for growth factor-induced gene expression and cell cycle entry. *EMBO J.* **18**, 664–674 (1999).
- Tsang, D. K. & Crowe, D. L. The mitogen activated protein kinase pathway is required for proliferation but not invasion of human squamous cell carcinoma lines. *Int. J. Oncol.* **15**, 519–523 (1999).
- Talarmin, H. *et al.* The mitogen-activated protein kinase/extracellular signal-regulated kinase cascade activation is a key signalling pathway involved in the regulation of G(1) phase progression in proliferating hepatocytes. *Mol. Cell. Biol.* **19**, 6003–6011 (1999).
- Klemke, R. L. *et al.* Regulation of cell motility by mitogen-activated protein kinase. *J. Cell Biol.* **137**, 481–492 (1997).
- Montesano, R., Soriano, J. V., Hosseini, G., Pepper, M. S. & Schramek, H. Constitutively active mitogen-activated protein kinase MEK1 disrupts morphogenesis and induces an invasive phenotype in Madin-Darby canine kidney epithelial cells. *Cell Growth Differ.* **10**, 317–332 (1999).
- Reddy, K. B., Krueger, J. S., Kondapaka, S. B. & Diglio, C. A. Mitogen-activated protein kinase (MAPK) regulates the expression of progelatinase B (MMP-9) in breast epithelial cells. *Int. J. Cancer* **82**, 268–273 (1999).
- Aguirre-Ghisio, J. A. *et al.* RalA requirement for v-Src- and v-Ras-induced tumorigenicity and

- overproduction of urokinase-type plasminogen activator: involvement of metalloproteases. *Oncogene* **18**, 4718–4725 (1999).
23. Esparza, J. *et al.* Fibronectin upregulates gelatinase B (MMP-9) and induces coordinated expression of gelatinase A (MMP-2) and its activator MT1-MMP (MMP-14) by human T lymphocyte cell lines. A process repressed through Ras/MAP kinase signaling pathways. *Blood* **94**, 2754–2766 (1999).
 24. Brett, J. *et al.* Tissue distribution of the receptor for advanced glycation endproducts (RAGE): expression in smooth muscle, cardiac myocytes, and neural tissue in addition to the vasculature. *Am. J. Pathol.* **143**, 1699–1712 (1993).
 25. O'Reilly, M. S. *et al.* Angiostatin: a novel angiogenesis inhibitor that mediates the suppression of metastases by a Lewis lung carcinoma. *Cell* **79**, 315–328 (1994).
 26. Leder, A., Kuo, A., Cardiff, R. D., Sinn, E. & Leder, P. V-Ha-ras transgene abrogates the initiation step in mouse skin tumorigenesis: effects of phorbol esters and retinoic acid. *Proc. Natl Acad. Sci. USA* **87**, 9178–9182 (1990).
 27. Yu, W., Kim, J. & Ossowski, L. Reduction in surface urokinase receptor forces malignant cells into a protracted state of dormancy. *J. Cell Biol.* **137**, 767–777 (1997).
 28. Valente, P. *et al.* TIMP-2 over-expression reduces invasion and angiogenesis and protects B16F10 melanoma cells from apoptosis. *Int. J. Cancer* **75**, 246–253 (1998).
 29. Albini, A. *et al.* A rapid *in vitro* assay for quantitating the invasive potential of tumor cells. *Cancer Res.* **47**, 3239–3245 (1987).
 30. Ramamurthy, N. S. & Golub, L. M. Diabetes increases collagenase activity in extracts of rat gingiva and skin. *J. Periodont. Res.* **18**, 23–30 (1983).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank P. D'Amore for advice and B. Tycko for assistance in these studies. This work was supported by the Surgical Research Fund of the College of Physicians & Surgeons, Columbia University, and by grants from the United States Public Health Service, Juvenile Diabetes Foundation International, and the American Heart Association, New York affiliate. G.D.T. is a recipient of a Faculty Development Award from the Robert Wood Johnson Foundation. A.M.S. is a recipient of a Burroughs Wellcome Fund Clinical Scientist Award in Translational Research.

Correspondence and requests for materials should be addressed to A.M.S. (e-mail: ams11@columbia.edu).

Neurotoxicity induces cleavage of p35 to p25 by calpain

Ming-sum Lee^{*†}, Young T. Kwon^{*†}, Mingwei Li[‡], Junmin Peng[†], Robert M. Friedlander[‡] & Li-Huei Tsai[†]

[†]Howard Hughes Medical Institute and Department of Pathology, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115, USA

[‡]Neuroapoptosis Laboratory, Neurosurgical Service, Department of Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA

* These authors contributed equally to this work

Cyclin-dependent kinase 5 (cdk5) and its neuron-specific activator p35 are required for neurite outgrowth and cortical lamination^{1–3}. Proteolytic cleavage of p35 produces p25, which accumulates in the brains of patients with Alzheimer's disease⁴. Conversion of p35 to p25 causes prolonged activation and mislocalization of cdk5. Consequently, the p25/cdk5 kinase hyperphosphorylates tau, disrupts the cytoskeleton and promotes the death (apoptosis) of primary neurons. Here we describe the mechanism of conversion of p35 to p25. In cultured primary cortical neurons, excitotoxins, hypoxic stress and calcium influx induce the production of p25. In fresh brain lysates, addition of calcium can stimulate cleavage of p35 to p25. Specific inhibitors of calpain, a calcium-dependent cysteine protease, effectively inhibit the calcium-induced cleavage of p35. *In vitro*, calpain directly cleaves p35 to release a fragment with relative molecular mass 25,000. The sequence of the calpain cleavage product corresponds precisely to that of p25. Application of the amyloid β -peptide A β (1–42) induces the conversion of p35 to p25 in primary cortical

neurons. Furthermore, inhibition of cdk5 or calpain activity reduces cell death in A β -treated cortical neurons. These observations indicate that cleavage of p35 to p25 by calpain may be involved in the pathogenesis of Alzheimer's disease.

The open reading frame of p35 does not contain introns², so alternative splicing cannot account for the generation of p25. Internal initiation of translation of p35 messenger RNA is also unlikely to produce p25 because there is no internal methionine near the beginning of the p25 sequence. Proteolytic cleavage is, therefore, the most likely mechanism for conversion of p35 to p25 (Fig. 1a).

Despite extensive efforts to identify p25 in the mouse, only full-length p35 was detectable during embryonic development and in the adult (data not shown). We next sought to determine whether p25 could be produced *in vivo* under certain experimental conditions. We found that 4 h of focal ischaemia, induced by middle cerebral artery occlusion in mice, produced p25 in the ipsilateral cortex but not in the control contralateral cortex (Fig. 1b). The conversion of p35 to p25 caused it to relocate to the cytoplasm (Fig. 1c), as reported previously⁴.

To investigate the mechanism of the conversion of p35 to p25 further, we tested for conditions that would induce the appearance of p25 in cultured primary cortical neurons. Treatment with hydrogen peroxide stimulated cleavage of p35 to p25 in primary neurons (Fig. 1d). Other insults such as treatment with the excitatory amino-acid glutamate also caused the production of p25 in cortical neurons at high concentrations of glutamate (Fig. 1e). An increase in intracellular calcium levels, caused by the calcium ionophore ionomycin, stimulated efficient conversion of p35 to p25 (Fig. 1f). These results indicate that neurotoxicity induces cleavage of p35 to p25 and suggest a role for calcium in this process.

To identify the protease that cleaves p35 to p25, we sought to recapitulate the proteolytic cleavage event. In fresh mouse brain lysates, 1 mM Ca²⁺ efficiently stimulates the cleavage of p35 (Fig. 2a). The cleavage product is likely to be p25, as it has a relative molecular mass of 25K and co-migrates with recombinant p25 expressed in COS-7 cells (lane 1). Also, like p25, it is specifically recognized by the p35 carboxy-terminal-specific antibody, but not by the p35 amino-terminal-specific antibody (see Supplementary Information).

p35 contains no obvious consensus sequences for cleavage by known proteases. To identify the protease activated by calcium, we tested protease inhibitors with different specificities for their effectiveness in inhibiting the calcium-stimulated p35 conversion. Calpeptin and calpain inhibitor II, which inhibit the calcium-dependent cysteine protease calpain, completely inhibited p35 cleavage (Fig. 2b, lanes 3–4), whereas the general cysteine protease inhibitor leupeptin partially inhibited p35 cleavage (lane 8). A titration of four calpain-specific inhibitors shows that 10 nM calpeptin, 100 nM calpain inhibitor I, 100 nM calpain inhibitor II and 5 nM calpastatin effectively inhibit p35 conversion (Fig. 2d and Supplementary Information), consistent with the reported median inhibitory concentration (IC₅₀) values for these inhibitors⁵. The lack of effect of the cdk5 inhibitor roscovitine indicates that cdk5 activity may not be necessary for cleavage to occur (Fig. 2b, lane 9).

m-calpain and μ -calpain are the two main isoforms of calpain in the brain⁶. The two calpains differ in their calcium requirements but have similar substrate specificities. μ -calpain requires 3–50 μ M calcium for half-maximal activity, whereas m-calpain requires 0.2–1 mM calcium for activity. To determine whether calpain was indeed activated in the conditions tested *in vitro*, we examined the cleavage of a well characterized calpain substrate, non-erythroid α -spectrin (also known as α -fodrin)⁵. One millimolar calcium, which stimulated conversion of p35 to p25 in mouse brain lysates, also led to cleavage of endogenous spectrin into the characteristic 145K and 150K fragments, indicating that calpain was activated (Fig. 2c). Furthermore, spectrin cleavage was inhibited by calpeptin,

calpain inhibitor I and leupeptin, showing that calpain activation correlates tightly with p35 cleavage.

To investigate further whether calpain cleaves p35 in brain lysates, we fractionated mouse brain lysates by glycerol gradient centrifugation and examined whether the p35 cleavage activity co-segregated with calpain activity. We incubated the glycerol gradient fractions with either purified p35 or spectrin; both p35 cleavage activity and calpain activity completely co-segregated in fractions 8–10 (Fig. 3a, b). Depletion of m-calpain from brain lysates using an m-calpain-specific antibody significantly reduced the Ca^{2+} -induced cleavage of p35 (Fig. 3c–e), further indicating that conversion of p35 to p25 may lie downstream of calpain activation.

To determine whether calpain directly cleaves p35, we incubated purified calpain with purified p35. Both m-calpain and μ -calpain cleaved p35 to a 25K fragment that co-migrates with p25 on SDS–PAGE (Fig. 3f). A titration of purified m-calpain and μ -calpain shows that 0.002 unit of m-calpain and 0.006 unit of μ -calpain cause half-maximal cleavage of p35 (Fig. 3g, h). To verify that the 25K fragment produced by calpain cleavage is indeed p25, we subjected p35 purified from a baculovirus expression system to calpain cleavage. The 25K calpain cleavage product was micro-sequenced from its N terminus. The first five amino acids from the N terminus of this fragment are AQPPP, which precisely matches the N terminus of p25 previously purified from bovine brain lysates⁷ (Fig. 3i).

To understand the mechanism for p35 cleavage in primary cortical neurons, we tested whether an increase in intracellular calcium is required for p35 conversion. H_2O_2 or ionomycin did not

stimulate conversion of p35 to p25 in neurons that were incubated in calcium-free medium (Fig. 4a). Similarly, removing calcium with either EGTA or BAPTA-AM also prevented conversion of p35. These results indicate that calcium is necessary for conversion of p35. The addition of calpeptin, which inhibits calpain activity in neurons (as indicated by the reduction of spectrin cleavage (Fig. 4c), also inhibited the cleavage of endogenous p35 to p25 caused by H_2O_2 (Fig. 4b). Similarly, conversion of p35 to p25 induced by ionomycin was completely reversed by increasing concentrations of calpeptin (Fig. 4d, e). Together, these experiments indicate that activation of calpain by neurotoxic processes through calcium influx is both necessary and sufficient for conversion of p35 to p25 in cortical neurons.

As p25 is found to accumulate in brains of Alzheimer's disease patients but not in normal brains, and because it is present in neurons containing neurofibrillary tangles⁴, we were interested in determining whether the amyloidogenic peptide $\text{A}\beta(1-42)$ can induce conversion of p35 to p25. The $\text{A}\beta$ peptides are primary constituents of the amyloid plaques found in the brains of patients with Alzheimer's disease. They have been shown to aggregate and cause neuronal death⁸. Treatment of primary cortical neurons with 20 μM $\text{A}\beta(1-42)$ peptide resulted in conversion of p35 to p25 and breakdown of endogenous spectrin from 280K to the 150K and 145K products (Fig. 5a, b). The same concentration of a control peptide, $\text{A}\beta(42-1)$, induced neither conversion of p35 to p25 nor cleavage of spectrin. These results indicate that, in Alzheimer's

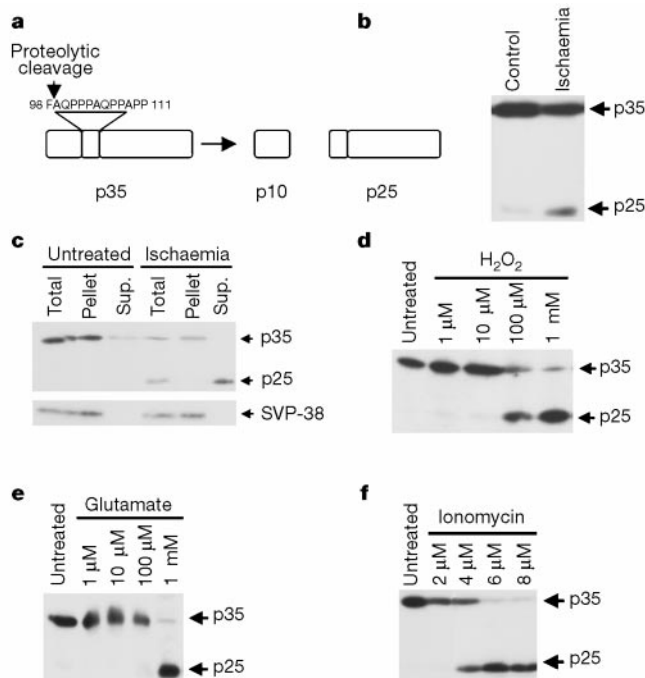


Figure 1 Neurotoxicity induces cleavage of p35 to p25. **a**, The cdk5 activator p35. Proteolytic cleavage of p35 between residues 98 and 99 liberates the neurotoxic 25K C-terminal fragment termed p25. **b**, Ischaemia induces conversion of p35 to p25. Western blot of p35 in the ipsilateral and control contralateral mouse brain tissue after 4 h of focal ischaemia. **c**, Distribution of p35 and p25 in fractionated brain lysates. The supernatant contains cytoplasmic proteins; the pellet corresponds to proteins associated with the membrane, cytoskeleton and nucleus. Synaptophysin (SVP-38) was used as a membrane marker. **d**, Oxidative stress induces p35 conversion. Neurons were treated with increasing amounts of H_2O_2 , lysed and assayed for p35 cleavage. **e**, Glutamate induces p35 conversion. Neuronal cultures were treated with increasing concentrations of glutamate and assayed for p35 cleavage. **f**, Ionomycin induces p35 conversion. Rat cortical neurons were treated with the indicated amount of ionomycin and assayed for p35 cleavage.

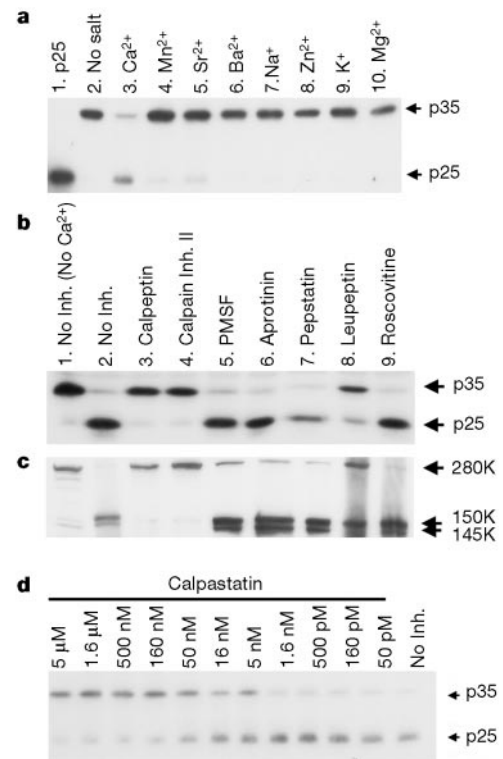


Figure 2 Ca^{2+} is necessary for p35 cleavage *in vitro*. **a**, Ca^{2+} induces p35 cleavage in mouse brain lysates. Western blot of p35 showing the effects of adding different chloride salts (1 mM) to mouse brain lysates. Lane 1, COS cell lysates transfected with recombinant p25 (control). **b**, Calpain inhibitors can inhibit conversion of p35 to p25. Calcium chloride was added to mouse brain lysate to stimulate p35 cleavage in the presence of the following inhibitors: 2 μM calpeptin (lane 3), 5 μM calpain inhibitor II (lane 4), 1 mM PMSF (lane 5), 1 μg μl^{-1} aprotinin (lane 6), 1 μg μl^{-1} pepstatin (lane 7), 1 μg μl^{-1} leupeptin (lane 8), 10 μM roscovitine (lane 9). **c**, Calpain activation correlates with p35 cleavage activity. Samples as in **b** were probed with an antibody against the endogenous calpain substrate non-erythroid α -spectrin. **d**, Mouse brain lysates treated with 5 mM calcium chloride and the indicated amount of calpastatin was assayed for p35 conversion.

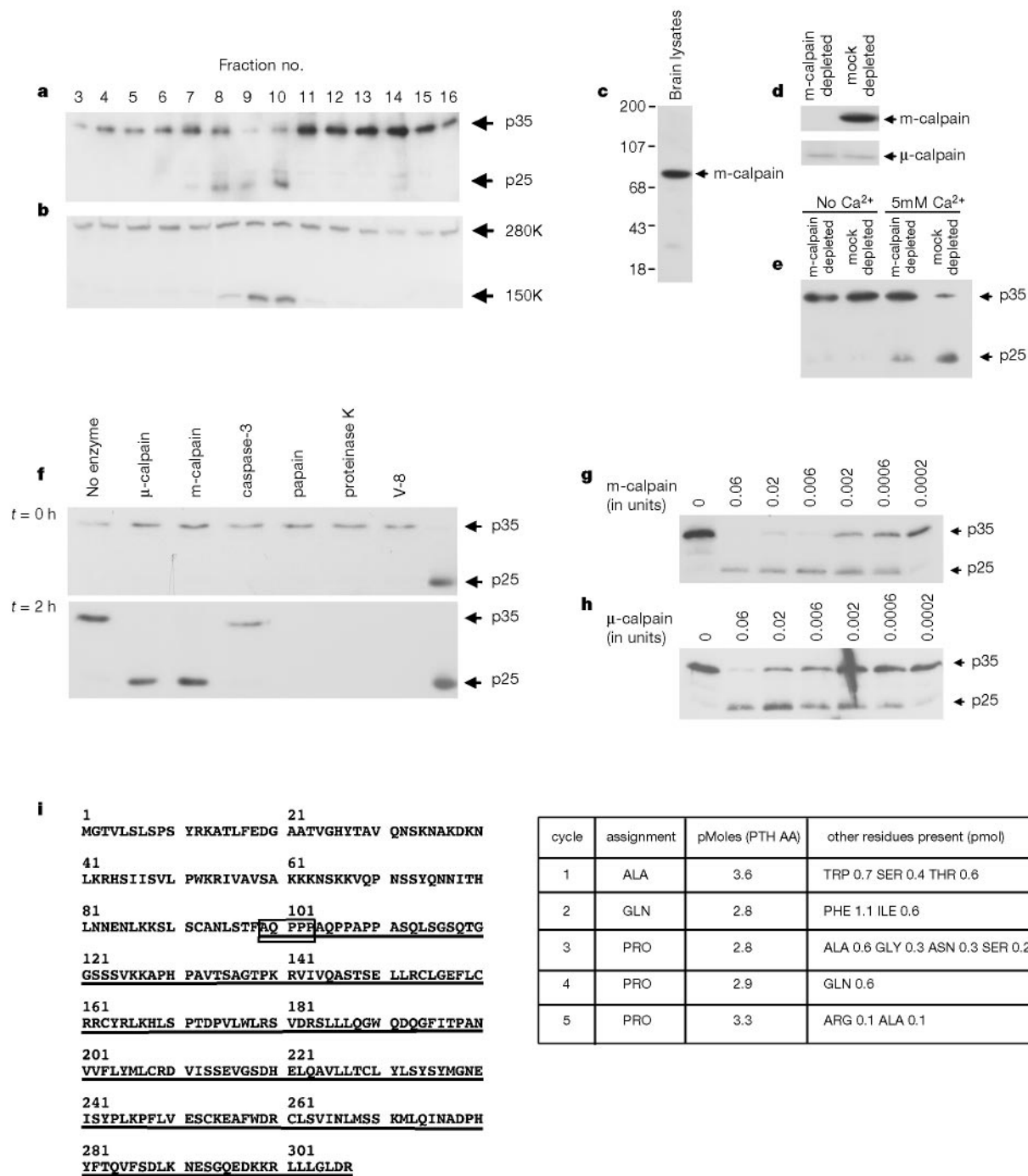


Figure 3 Calpain directly cleaves p35 to p25. **a**, Calpain activity co-fractionates with p35-cleaving activity. Mouse brain lysate was fractionated by a glycerol gradient. Each fraction was incubated with p35 immunoprecipitated from mouse brain lysates for 2 h at 25 °C to assay for p35 cleavage activity. **b**, The same fractions as in **a** were incubated with inactivated mouse brain lysate (as a source of spectrin) and assayed for spectrin cleavage activity. **c**, Western blot on rat brain lysates showing specificity of the m-calpain antibody. **d**, Western blots showing the levels of m-calpain and μ -calpain in rat brain lysates after immunodepletion. Endogenous m-calpain was completely depleted, whereas μ -calpain levels remained the same. **e**, Immunodepleting m-calpain from rat brain lysates substantially reduces the p35-cleavage activity. Rat brain lysates depleted 3 times with the m-calpain specific antibody was treated with 5 mM Ca^{2+} and assayed for p35

cleavage. **f**, Calpain cleaves p35. p35 immunoprecipitated from mouse brain lysates was treated with proteases as indicated. Western blot of p35 before (upper panel) and after (lower panel) protease treatment. **g**, Decreasing concentrations of m-calpain was incubated with p35 purified from baculovirus-infected insect cells for 30 min at 30 °C and assayed for p35 cleavage. **h**, as in **g**, but with μ -calpain. **i**, The calpain cleavage product has the same sequence as p25. The entire sequence of p35 is shown, the sequence for p25 is underlined. Recombinant p35 from baculovirus-infected insect cells was purified and treated with m-calpain. The calpain cleavage product was microsequenced. The boxed area shows the first five N-terminal residues of the cleavage product obtained from protein microsequencing. The table shows the concentrations of residues present in each sequencing cycle in pmol.

cycle	assignment	pMoles (PTH AA)	other residues present (pmol)
1	ALA	3.6	TRP 0.7 SER 0.4 THR 0.6
2	GLN	2.8	PHE 1.1 ILE 0.6
3	PRO	2.8	ALA 0.6 GLY 0.3 ASN 0.3 SER 0.2
4	PRO	2.9	GLN 0.6
5	PRO	3.3	ARG 0.1 ALA 0.1

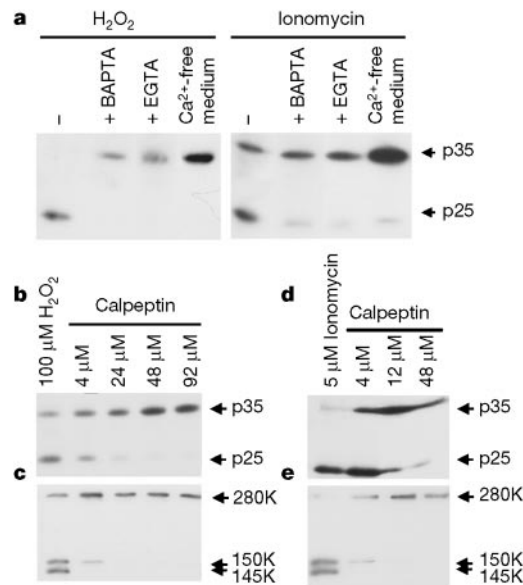


Figure 4 Conversion of p35 to p25 in primary cortical neurons is mediated by calpain and requires Ca^{2+} . **a**, Left panel, primary cortical neurons were treated with 1 mM H_2O_2 and 30 μM BAPTA-AM, 5 mM EGTA or phosphate-buffered saline (PBS). Right panel, neurons treated with 4 μM ionomycin were incubated with 500 μM BAPTA-AM, 5 mM EGTA or PBS and assayed for p35 conversion. **b**, H_2O_2 -induced p35 conversion is mediated by calpain. 100 μM H_2O_2 was added to neurons treated with increasing amounts of the

calpain inhibitor calpeptin, and the samples were assayed for p35 cleavage. **c**, as in **b**, but samples were assayed for cleavage of spectrin. **d**, The calpain inhibitor calpeptin can inhibit ionomycin-stimulated conversion of p35 to p25. 5 μM ionomycin was added to cultures treated with increasing amounts of calpeptin, as indicated, and assayed for p35 cleavage. **e**, as in **d**, but cleavage of spectrin was assayed.

disease, the abundant amyloid plaques in the brain can cause calpain activation and conversion of p35 to p25.

$\text{A}\beta(1-42)$ induces cell death in cell-culture models. As $\text{A}\beta(1-42)$ causes conversion of p35 to p25, and the p25/cdk5 kinase induces cell death in cultured primary cortical neurons⁴, we were interested in determining whether the p25/cdk5 kinase is involved in $\text{A}\beta(1-42)$ peptide-induced cell death. Cells treated with 10 μM $\text{A}\beta(1-42)$ showed markedly reduced survival compared with mock-treated neurons (Fig. 5c). We found that 10 μM of the cdk5 inhibitor butyrolactone significantly inhibited neuronal cell death caused by $\text{A}\beta(1-42)$. Similarly, 20 μM of the calpain inhibitor calpeptin considerably reduced $\text{A}\beta(1-42)$ -induced neuronal death. Inhibiting cdk5 activity by expressing a cdk5 antisense construct in hippocampal neurons also inhibits cell death caused by $\text{A}\beta(1-40)$ ⁹. Together, these results indicate that active cdk5 is involved in $\text{A}\beta$ -induced cell death.

The tight regulation of p35 allows cdk5 to be activated in a

temporally and spatially specific manner. In contrast, p25 causes mislocalization and constitutive activation of cdk5. The cleavage of p35 to p25 by calpain represents a new regulatory mechanism whereby proteolytic cleavage deregulates a protein kinase. By cleaving p35, calpain does not destroy cdk5 activity; rather, it alters the properties of cdk5 such that the p25/cdk5 kinase causes collapse of the cytoskeleton and cell death. Thus, whereas p35/cdk5 activity is necessary for proper development and other functions of the mature nervous system, p25/cdk5 activity is detrimental to neurons.

Altered calcium homeostasis is a convergence point of various aetiological factors relevant to the pathogenesis of Alzheimer's disease. In particular, loss of calcium homeostasis has been implicated in causing tau hyperphosphorylation and neuronal apoptosis¹⁰⁻¹³. This study, showing cleavage of p35 to p25 by calpain, suggests one mechanism by which calcium can activate a tau-phosphorylating kinase. Consistent with our model, activated m-calpain and μ -calpain are both abnormally upregulated in the

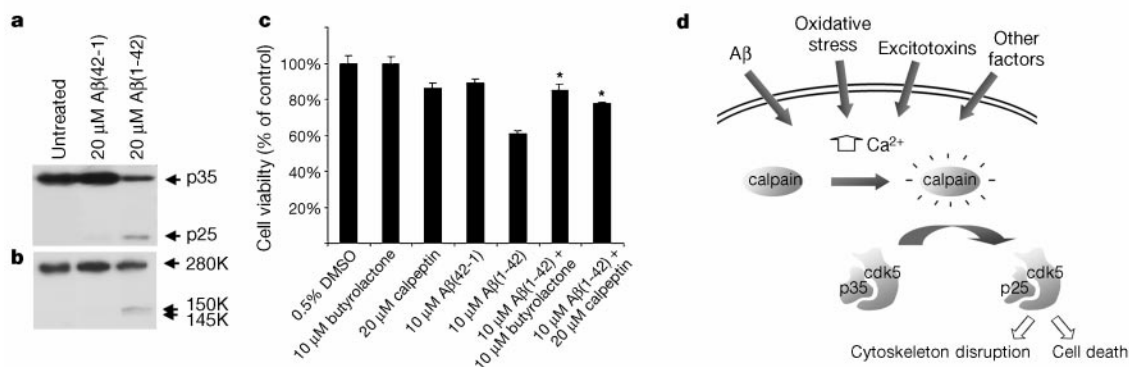


Figure 5 $\text{A}\beta(1-42)$ induces conversion of p35 to p25; inhibition of cdk5 or calpain reduces $\text{A}\beta(1-42)$ -induced cell death. **a**, $\text{A}\beta(1-42)$ causes generation of p25. Primary cortical neurons treated with 20 μM of $\text{A}\beta(1-42)$ or $\text{A}\beta(42-1)$ for 4 days were assayed for p35 cleavage. **b**, as in **a**, but samples were assayed for spectrin cleavage. **c**, Inhibition of $\text{A}\beta(1-42)$ -induced cell death. Four-day-old primary cortical neurons were treated with

either 0.5% DMSO, 10 μM butyrolactone, 20 μM calpeptin, 10 μM $\text{A}\beta(1-42)$ or 10 μM $\text{A}\beta(42-1)$. Cell viability after a 24-h treatment was assayed with MTT and compared with untreated neurons. Data are averages ($\pm$ s.e.) of four independent experiments. Asterisk, $P < 0.01$ compared with $\text{A}\beta(1-42)$ treated neurons. **d**, Neurotoxic processes activate the protease calpain, which results in p25 production and deregulation of cdk5.

brains of patients with Alzheimer's disease^{14–17}. In addition, active m-calpain accumulates in neurofibrillary tangles in Alzheimer's disease brains¹⁴.

The amyloid hypothesis proposes that the A β peptides, perhaps acting with other influences such as excitotoxins, free radicals and oxidative stress, cause an increased or unregulated entry of calcium into affected cells, which ultimately leads to the activation of a tau-phosphorylating kinase¹⁸. Our observations indicate that cdk5 may be one of the kinases 'activated' by amyloid- β peptide through calpain-mediated conversion of p35 to p25 (Fig. 5d). Given the potentially deleterious role of cdk5 in Alzheimer's disease, the calpain-mediated p35 cleavage pathway may serve as a target for pharmacological intervention. $\square$

Methods

Chemicals and antibodies

p25 antibody was raised against the whole protein and purified against glutathione S-transferase (GST)-p25. Non-erythroid α -spectrin antibody was purchased from Chemicon. p35 C-19, p35 N-20 and μ -calpain polyclonal antibodies were purchased from Santa-Cruz. The anti-rat m-calpain antibody is a gift from J. Elce. PMSF, Pepstatin A, aprotinin, leupeptin, MTT, glutamate and all metal chlorides were purchased from Sigma. Purified calpain I, purified calpain II, calpastatin, calpeptin, calpain inhibitor I, calpain inhibitor II, BAPTA-AM, butyrolactone and ionomycin were purchased from Calbiochem. H₂O₂ was purchased from Fisher Scientific. A β (1–42) and A β (42–1) were purchased from Bachem.

Primary cortical neuronal cultures

E17–E19 pregnant Long Evans rats were purchased from Harland Sprague–Dawley. Embryos were surgically removed and their cortices were dissected and cultured as described³. Cortical cultures were grown in basal growth medium on 6-well plates coated with laminin and poly-D-lysine. Treatments with H₂O₂ and glutamate were performed 10 d after plating for 5 h. Treatment with ionomycin was performed 4 d after plating for 5 h. In experiments where calpeptin, EGTA, or BAPTA-AM were used, the drugs were added 30 min before challenges were applied.

Glycerol gradient

A 11-ml 10–25% glycerol gradient was set up in ELB buffer (50 mM Tris pH 7, 0.1% NP-40, 250 mM NaCl). We layered 300 μ l of fresh mouse brain lysate on top of the gradient and centrifuged at 40K r.p.m. for 26 h. We collected 17 600- μ l fractions. We incubated 10 μ l of each fraction at 25 °C for 2 h with either purified p35 or frozen and thawed mouse brain lysates in a reaction buffer (5 mM CaCl₂, 5 mM cysteine, 150 mM imidazole, pH 7.5) to a final volume of 100 μ l. The product from each reaction was electrophoresed on a 12% gel and analysed as described in the text.

Protein microsequencing

Recombinant p35 and His-tagged cdk5 were produced by baculovirus and the p35/cdk5 complex was purified through a Ni²⁺ column. We treated 1 μ g of the purified p35/cdk5 complex with 2 units of purified m-calpain (Calbiochem) at 30 °C for 30 min. The reaction was stopped by 1% SDS. The entire reaction was electrophoresed on a 15% acrylamide gel and transferred to a 0.2- μ m PVDF membrane (BioRad). The membrane was stained with Coomassie Brilliant Blue (BioRad), the 25K cleavage product was excised, and 1/3 of the excised band was sequenced on Applied Biosystems Model 494 Procise Protein Sequencer with Model 140C Microgradient Delivery System and Model 785A Programmable Absorbance Detector.

Ischaemia

To induce focal ischaemia, adult mice (C57BL/6), weighing 16–20 g, were anaesthetized initially with 1.5% isoflurane and thereafter maintained in 1.0% isoflurane in 70% N₂O and 30% O₂. An 8.0 nylon monofilament suture coated with a silicone/hardener mixture (Heraeus Kulzer) was inserted into the right common carotid artery. The suture was advanced 9–10 mm from the insertion site through the internal carotid artery, occluding the middle cerebral artery (MCA). Mice all woke up hemiplegic and were killed with isoflurane 4 h after the induction of ischaemia. Global ischaemia was induced for 1 h as described¹⁹. Generation and fractionation of brain lysates was as described²⁰.

Cell viability assay

Primary cortical neurons were plated on 96-well plates. Four days after plating, neurons were incubated with 20 μ M A β (1–42) (from a stock of 1 mg ml^{–1} in ddH₂O) with or without 10 μ M butyrolactone (from 2.8 mM stock in DMSO) or 20 μ M calpeptin (from 2 mM stock in DMSO) for 24 h. Cell viability was measured by treating neurons with MTT (Sigma) at a final concentration of 0.5 mg ml^{–1} for 1 h. After MTT treatment, cell medium was replaced with 100 μ l DMSO and the optical density of each well at 495 nm was determined using a microtitre plate reader.

Received 16 December 1999; accepted 7 April 2000.

- Ohshima, T. *et al.* Targeted disruption of the cyclin-dependent kinase 5 gene results in abnormal corticogenesis, neuronal pathology and perinatal death. *Proc. Natl Acad. Sci. USA* **93**, 11173–11178 (1996).

- Chae, T. *et al.* Mice lacking p35, a neuronal specific activator of Cdk5, display cortical lamination defects, seizures, and adult lethality. *Neuron* **18**, 29–42 (1997).
- Nikolic, M., Dudek, H., Kwon, Y. T., Ramos, Y. F. & Tsai, L. H. The cdk5/p35 kinase is essential for neurite outgrowth during neuronal differentiation. *Genes Dev.* **10**, 816–825 (1996).
- Patrick, G. N. *et al.* Conversion of p35 to p25 de-regulates cdk5 activity and promotes neurodegeneration. *Nature* **402**, 615–622 (1999).
- Wang, K. K. W. in *Calpain: Pharmacology and Toxicology of Calcium-Dependent Protease* (eds Wang, K. K. W. & Yuen, P.-W.) 77–102 (Taylor & Francis, Philadelphia, 1999).
- Sorimachi, H., Ishiura, S. & Suzuki, K. Structure and physiological function of calpains. *Biochem. J.* **328**, 721–732 (1997).
- Lew, J. *et al.* A brain-specific activator of cyclin-dependent kinase 5. *Nature* **371**, 423–426 (1994).
- Yankner, B. A. Mechanisms of neuronal degeneration in Alzheimer's disease. *Neuron* **16**, 921–932 (1996).
- Alvarez, A., Toro, R., Caceres, A. & Maccioni, R. B. Inhibition of tau phosphorylating protein kinase cdk5 prevents β -amyloid-induced neuronal death. *FEBS Lett.* **459**, 421–426 (1999).
- Arispe, N., Pollard, H. B. & Rojas, E. Giant multilevel cation channels formed by Alzheimer disease amyloid beta protein [A β P-(1–40)] in bilayer membranes. *Proc. Natl Acad. Sci. USA* **90**, 10573–10577 (1993).
- Hartmann, H., Eckert, A. & Muller, W. E. β -Amyloid protein amplifies calcium signalling in central neurons from the adult mouse. *Biochem. Biophys. Res. Commun.* **194**, 1216–1220 (1993).
- Mattson, M. P. *et al.* β -Amyloid peptides destabilize calcium homeostasis and render human cortical neurons vulnerable to excitotoxicity. *J. Neurosci.* **12**, 376–389 (1992).
- Pascale, A. & Etcheberrygaray, R. Calcium alterations in Alzheimer's disease: pathophysiology, models and therapeutic opportunities. *Pharmacol. Res.* **39**, 81–88 (1999).
- Grynspar, R., Griffin, W. R., Cataldo, A., Katayama, S. & Nixon, R. A. Active site-directed antibodies identify calpain II as an early-appearing and pervasive component of neurofibrillary pathology in Alzheimer's disease. *Brain Res.* **763**, 145–158 (1997).
- Nixon, R. A. *et al.* Calcium-activated neutral proteinase (calpain) system in aging and Alzheimer's disease. *Ann. NY Acad. Sci.* **747**, 77–91 (1994).
- Saito, K., Elce, J. S., Hamos, J. E. & Nixon, R. A. Widespread activation of calcium-activated neutral proteinase (calpain) in the brain in Alzheimer disease: a potential molecular basis for neuronal degeneration. *Proc. Natl Acad. Sci. USA* **90**, 2628–2632 (1993).
- Iwamoto, N., Thangnipon, W., Crawford, C. & Emson, P. C. Localization of calpain immunoreactivity in senile plaques and in neurones undergoing neurofibrillary degeneration in Alzheimer's disease. *Brain Res.* **561**, 177–180 (1991).
- Selkoe, D. J. Translating cell biology into therapeutic advances in Alzheimer's disease. *Nature* **399**, A23–A31 (1999).
- Seisjo, B. K. in *Brain Energy Metabolism* 454 (Wiley, New York, 1978).
- Nikolic, M., Chou, M. M., Lu, W., Mayer, B. J. & Tsai, L. H. The p35/Cdk5 kinase is a neuron-specific Rac effector that inhibits Pak1 activity. *Nature* **395**, 194–198 (1998).

Acknowledgements

We thank J. Elce for the m-calpain antibody; A. Bonni and V. Tannoch for reading this manuscript; M. Ahljanian, B. Yankner, E. Wang, K. Kosik and members of the Tsai lab for discussions. This work was partially supported by NIH grants to L.-H.T. L.-H.T. is an assistant investigator of the Howard Hughes Medical Institute, a Rita Allen Foundation scholar and a recipient of an Ester A. and Joseph Kligenstein Fund grant.

Correspondence and requests for materials should be addressed to L.-H.T. (e-mail: li-huei_tsai@hms.harvard.edu).

Supplementary information is available from Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

LAG-3 is a putative transcriptional activator in the *C. elegans* Notch pathway

Andrei G. Petcherski & Judith Kimble

Howard Hughes Medical Institute and Department of Biochemistry, University of Wisconsin-Madison, 433 Babcock Drive, Madison, Wisconsin 53706, USA

Notch signalling controls growth, differentiation and patterning during normal animal development^{1,2}; in humans, aberrant Notch signalling has been implicated in cancer and stroke^{3,4}. The mechanism of Notch signalling is thought to require cleavage of the receptor in response to ligand binding⁵, movement of the receptor's intracellular domain to the nucleus^{6,7}, and binding of that intracellular domain to a CSL (for CBF1, Suppressor of

Hairless, LAG-1)^{8,9} protein. Here we identify LAG-3, a glutamine-rich protein that forms a ternary complex together with the LAG-1 DNA-binding protein¹⁰ and the receptor's intracellular domain. Receptors with mutant ankyrin repeats that abrogate signal transduction are incapable of complex formation both in yeast and *in vitro*. Using RNA interference, we find that LAG-3 activity is crucial in *Caenorhabditis elegans* for both GLP-1 and LIN-12 signalling. LAG-3 is a potent transcriptional activator in yeast, and a Myc-tagged LAG-3 is predominantly nuclear in *C. elegans*. We propose that GLP-1 and LIN-12 promote signalling by recruiting LAG-3 to target promoters, where it functions as a transcriptional activator.

The *C. elegans* Notch signalling pathway relies on either of two receptors, GLP-1 or LIN-12 (Fig. 1a)^{1,2}. The intracellular portions of GLP-1 and LIN-12 possess three distinct regions: the RAM domain binds tightly and specifically to LAG-1 (ref. 8); the ankyrin repeats (ANK) are critical for signalling^{11–13}; and the PEST domain is thought to control receptor stability¹⁴ (Fig. 1b). Among these domains, the RAM and ANK regions are sufficient for eliciting a strong signalling response⁸.

We used a modified yeast two-hybrid screen to identify proteins that interact with the GLP-1 intracellular domain only in the presence of LAG-1. Briefly, one hybrid protein, GLP-1(RAM-ANK), fused to the LexA DNA-binding domain and one non-hybrid protein, LAG-1, were used as a complex bait, selecting for activation of the *HIS3* gene in the presence of a *C. elegans* complementary DNA/activation domain fusion library (see Methods). Out of 300,000 transformants, 6 were dependent on the presence of LAG-1 for robust growth; all 6 positives carried a complementary DNA encoding the same glutamine-rich protein (Fig. 1c), which we call LAG-3. In yeast, LAG-3 interacts strongly with both GLP-1(RAM-ANK) and LIN-12(RAM-ANK) in the presence of LAG-1, but not in its absence (Fig. 1d). Furthermore, LAG-3 does not interact with either LAG-1 alone or a lamin control (Fig. 1d). The simplest interpretation is that a ternary complex is formed which is composed of LAG-3, LAG-1 and either GLP-1(RAM-ANK) or LIN-12(RAM-ANK).

We then assayed formation of the complex *in vitro*. We generated

Table 1 Wild-type ANK repeats are required in yeast for complex formation and LAG-3 is a potent transcriptional activator in yeast

AD-Lag-3	T (°C)	GLP-1(RAM-ANK)*			LIN-12(RAM-ANK)*		LAG-1*	Lamin*
		wt	q224	q231	wt	n653		
Lag-1	15	++++	++++	++++	++++	–	±	–
	20	++++	++	++++	++++	–	–	–
	25	++++	+	+++	+++	–	–	–
	30	++++	+	++	++	–	–	–
no Lag-1	15	++	–	–	–	–	±	–
	20	++	–	–	–	–	–	–
	25	++	–	–	–	–	–	–
	30	++	–	–	–	–	±	–

* LexA fusion proteins were used. T, temperature of yeast growth. Scale for β -galactosidase activity: +, +++, strong blue colour detected by 30 min; ++, by 60 min; +, by 90 min; ±, by 120 min; –, no blue detected at 4 h.

a LAG-1 carrying a T7 epitope tag (Novagen) bound to beads and versions of the remaining components, each bearing an S-peptide tag (Novagen). After incubation, the proteins associating with beads were detected by western blot (see Methods). LAG-3 bound LAG-1 in the presence of wild-type GLP-1(RAM-ANK) (Fig. 2a, lane 3), but did not bind LAG-1 alone (Fig. 2a, lane 7). As expected from previous results⁸, GLP-1(RAM-ANK) bound LAG-1 regardless of LAG-3 (Fig. 2a, lane 5). To determine whether LAG-3 binds GLP-1(RAM-ANK) on its own, we carried out a reciprocal experiment with T7-LAG-3 bound to beads and S-tagged versions of LAG-1 and GLP-1(RAM-ANK). Wild-type GLP-1(RAM-ANK) bound weakly to LAG-3 (Fig. 2b, lanes 7, 8), but binding was enhanced markedly by the addition of LAG-1 (Fig. 2b, lanes 3, 4). We conclude that a ternary complex composed of LAG-1, LAG-3 and the RAM-ANK region of GLP-1 forms *in vitro*.

Previous work has shown that the ANK repeats of both GLP-1 and LIN-12 are essential for signalling^{11–13}. To determine whether ANK repeat mutations disrupted ternary complex formation, we assayed three mutations in the fourth ANK repeat of either GLP-1 or

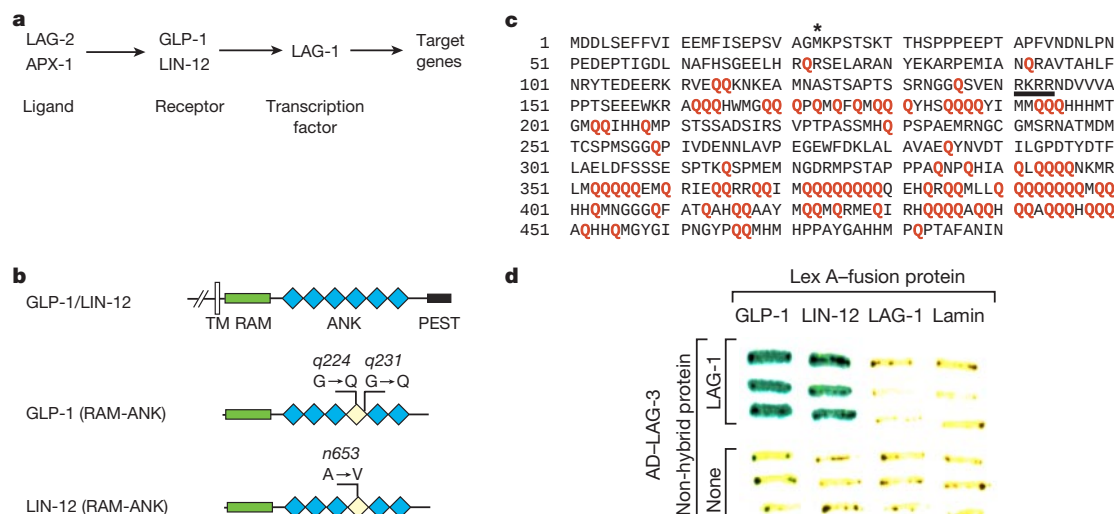


Figure 1 LAG-3 and ternary complex formation in yeast. **a**, Core components of the GLP-1/LIN-12 pathway (reviewed in ref. 1). **b**, Intracellular domain of receptors. Top, wild type proteins. TM, transmembrane domain; RAM, RAM domain (green box); ANK, ankyrin repeats (blue diamonds); PEST, PEST domain (black box). Middle, GLP-1 mutant protein. Bottom, LIN-12 mutant protein. **c**, LAG-3 amino-acid sequence. Glutamine is highlighted in red; the predicted nuclear localization sequence is underlined. Two alternative 5' ends

predict two proteins, LAG-3A and LAG-3B (see Methods). Asterisk denotes the methionine predicted to initiate LAG-3B. **d**, Interactions of LAG-1, LAG-3 and GLP-1(RAM-ANK) or LIN-12(RAM-ANK) proteins in yeast. Transcription by LexA-controlled reporter yields β -galactosidase (blue). Top, LexA fusion proteins; side, AD-LAG-3, GAL-4-activation domain fused to LAG-3. Yeast were grown at 20 °C.

LIN-12 (Fig. 1b); all three mutant proteins were produced stably in yeast as shown by western blot (not shown). In *C. elegans*, the *glp-1(q224)* and *glp-1(q231)* mutant receptors signal almost normally at 15 °C, but function poorly at 20 °C and 25 °C (ref. 15). When assayed in yeast, both mutants were defective for ternary complex formation in a temperature-sensitive manner: interactions were indistinguishable from wild-type at 15 °C, but much weaker at 30 °C (Table 1). Moreover, the strength of interaction correlated with allelic strength: the weaker *q231* mutant interacted better in yeast than did the more severe *q224* mutant. Similarly for *lin-12*, the *lin-12(n653)* mutant receptor is severely impaired in *C. elegans*¹², and when assayed in yeast, LIN-12(RAM-ANKn653) did not participate in complex formation (Table 1). Finally, when the mutant GLP-1(RAM-ANKq224) protein was assayed *in vitro*, formation of the ternary complex was not detectable (Fig. 2a, lane 4; Fig. 2b, lanes 5, 6). We conclude that the ANK repeats of both GLP-1 and LIN-12 receptors are critical for interactions with LAG-1 and LAG-3, and suggest that the failure of the mutant receptors to function properly in *C. elegans* may be due to their inability to form ternary complexes.

We then determined whether *lag-3* activity is required for *glp-1* and *lin-12* signalling in *C. elegans*. We used RNA interference (RNAi)¹⁶ to reduce *lag-3* activity (see Methods). To examine the effects of *lag-3(RNAi)* on early embryos, mothers were soaked in double-stranded (ds) *lag-3* RNA as late L4 larvae. Whereas wild-type embryos have an elongate pharynx (Fig. 3a), both *glp-1(q224)* and *lag-3(RNAi)* embryos lack an anterior pharynx and die within the eggshell (Fig. 3b,c). To examine the effects of *lag-3(RNAi)* postembryonically, early L1 larvae were soaked in *lag-3* dsRNA. Resultant *lag-3(RNAi)* adults were usually sterile (91%, *n* = 100), and sometimes possessed a protruding vulva (36%, *n* = 100). The sterility of *lag-3(RNAi)* adults resulted from a failure in germline proliferation (Fig. 3d,e), a phenotype typical of *glp-1* null mutants¹⁵; the protruding vulva (Fig. 3f) was similar to that of *lin-12* loss-of-function mutants¹⁷. More diagnostic of a *lin-12* signalling defect, however, was the finding of two anchor cells in

lag-3(RNAi) larvae (Fig. 3g). Also, these *lin-12* and *glp-1* defects were the only ones observed and, therefore, other pathways appear to be unaffected. From these results, we conclude that *lag-3* is critical for *glp-1* signalling in both embryo and germ line and for *lin-12* signalling postembryonically.

To further investigate the role of *lag-3* in *C. elegans*, we assayed the effect of *lag-3(RNAi)* on both *glp-1* and *lin-12* gain-of-function mutants. Whereas most *glp-1(oz112 gf)* animals are sterile and make tumorous germ lines at 25 °C (ref. 18) (Fig. 3h, upper row), most *glp-1(oz112 gf); lag-3(RNAi)* adults are either fertile or sterile with a Glp-1-like germ line composed solely of a few mature sperm (Fig. 3h, lower row). Thus, *lag-3* is critical for *glp-1(gf)* signalling. Similarly for *lin-12*, all *lin-12(n137 gf)* mutants possess multiple pseudovulvae, the Muv phenotype (Fig. 3i, upper row), and do not possess an anchor cell¹⁷, whereas most *lin-12(n137 gf); lag-3(RNAi)* animals have either a normal or a Lin-12-like protruding vulva (Fig. 3i, lower row). Upon examination by DIC microscopy, most

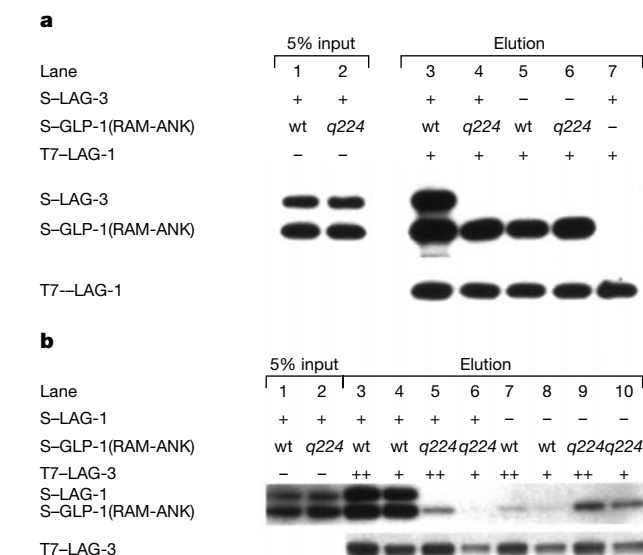


Figure 2 LAG-3 forms a ternary complex *in vitro* with GLP-1(RAM-ANK) and LAG-1. **a**, LAG-3 binds LAG-1 in the presence of wild-type GLP-1(RAM-ANK). Lanes 1, 2, input S-tagged proteins; lane 3–10, T7-LAG-1 was bound to beads; proteins incubated with beads shown at top of each lane. **b**, LAG-1 binds LAG-3 in presence of wild-type GLP-1. Lanes 1, 2, input S-tagged proteins; lanes 3–10, T7-LAG-3 was bound to beads; proteins incubated with beads shown at top of each lane. Assays done in duplicate using two different amounts of LAG-3 for each experiment (++ indicates that the amount of T7-LAG-3 is ~3 times greater than +).

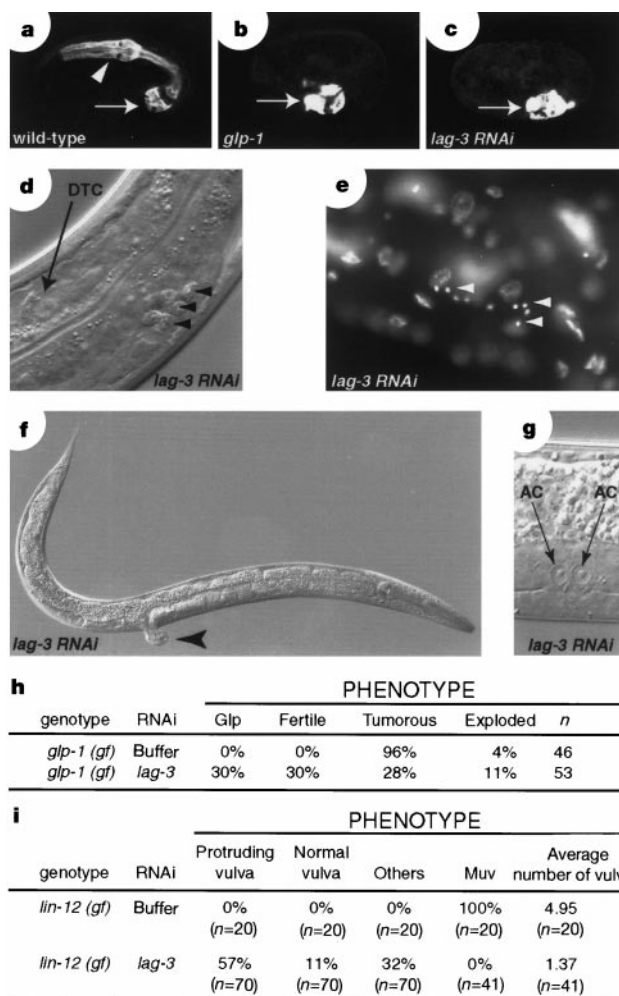


Figure 3 *lag-3* is critical for both *glp-1* and *lin-12* signalling. **a–c**, Embryos stained with 3NB12 antibody: arrowhead, anterior pharynx; arrow, posterior pharynx. **d–e**, *lag-3(RNAi)* adults with germ line composed solely of a few mature sperm (arrowheads). **d**, Distal tip cell (DTC). **e**, 4',6-diamidino-2-phenylindole (DAPI) stained. **f**, *lag-3(RNAi)* adult with protruding vulva (arrowhead). **g**, *lag-3(RNAi)* L3 with two anchor cells (AC) (arrows). **h–i**, Suppression of *glp-1(oz112 gf)* and *lin-12(n137 gf)* by *lag-3(RNAi)*. L1s soaked with either *lag-3* dsRNA or M9 buffer. *n*, number of animals. **h**, Complete genotype: *unc-32(e189) glp-1(oz112 gf)*. Glp, *glp-1(lf)* phenotype. Tumorous, *glp-1(gf)* phenotype. Exploded indicates that animals could not be scored. **i**, Suppression of *lin-12(n137 gf)* by *lag-3(RNAi)*. Protruding vulva, *lin-12(lf)* phenotype; others, minor vulval defects. Muv, *lin-12(gf)* phenotype.

lin-12(gf); lag-3(RNAi) animals possessed at least one anchor cell (data not shown). Both *glp-1(gf)* and *lin-12(gf)* receptors therefore rely on *lag-3* activity for signalling. We conclude that *lag-3* is an integral component of both *glp-1* and *lin-12* signalling pathways.

How might LAG-3 function in signal transduction? The LAG-3 amino-acid sequence is glutamine-rich, containing 20% glutamine residues (Fig. 1c), a hallmark of certain transcriptional activation domains¹⁹. To determine whether LAG-3 might function in transcriptional control, we did two experiments. First, we tested a LexA-LAG-3 fusion protein for its ability to activate transcription in yeast, and found it to be a strong transcriptional activator (Table 1). In contrast, other nuclear components of the *glp-1/lin-12* pathway possessed either no activity by this assay or a much more modest activity (Table 1). Second, we examined the subcellular localization of a Myc-tagged LAG-3 fusion protein placed under heat shock control and introduced into transgenic animals. Upon heat shock, Myc-LAG-3 was predominantly nuclear (Fig. 4a), whereas without heat shock, virtually no staining could be detected (Fig. 4b). Consistent with the idea that LAG-3 is a nuclear protein, there is a classical nuclear localization signal²⁰ in its amino-acid sequence (Fig. 1c, underline). The simplest interpretation is that LAG-3 normally serves as a transcriptional activator and that the primary function of the receptor's intracellular domain is to bring LAG-3 and LAG-1 together in a ternary complex at the target gene promoter.

Figure 4c,d illustrates one model for how LAG-3 and ternary complex formation might function in *glp-1* and *lin-12* signalling. In

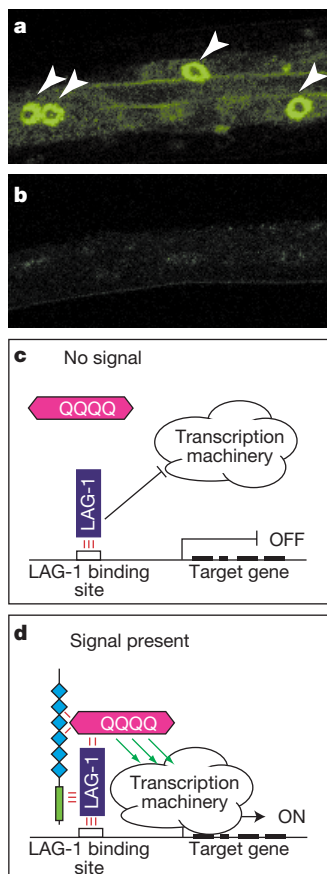


Figure 4 LAG-3 may function as a transcriptional activator. **a,b**, Adults carrying Myc-LAG-3 under heat shock control, examined using anti-Myc antibodies. **a**, After heat shock, Myc-LAG-3 is predominantly nuclear, as seen in intestinal nuclei (arrowheads). **b**, Without heat shock, Myc-LAG-3 is barely detectable. **c,d**, Model for LAG-3 function. **c**, Without signalling, no ternary complex is formed at target gene promoter and transcription is repressed. **d**, Upon signalling, the ternary complex forms and transcription is activated.

the absence of ligand binding (Fig. 4c), the receptor's intracellular domain does not enter the nucleus and LAG-3 is not recruited to target gene promoters. Upon ligand binding (Fig. 4d), however, the receptor is cleaved and its intracellular domain moves to the nucleus; as a result, LAG-3 is recruited to the promoter as part of a ternary complex, activating transcription. It must be emphasized that this model is speculative. For example, both the specific binding interactions within the ternary complex and the stoichiometry of components within the complex are not known. If LAG-3 is indeed a transcriptional activator in *C. elegans*, which we think is likely, the LAG-3 glutamine-rich domain might serve as the primary activation domain, as illustrated, or it might act together with other proteins, for example with the receptor's intracellular domain.

The identification of a new component in the *C. elegans glp-1/lin-12* signalling pathway has important implications for our understanding of Notch signalling. In particular, evidence suggests that a *lag-3*-like activity will be found in *Drosophila* and vertebrates. First, the strong conservation among CSL DNA-binding proteins and the conservation of both RAM and ANK repeat regions in all Notch-like receptors suggests that these proteins participate in a conserved process. Second, missense mutations in the ANK repeats of *Drosophila* Notch²¹ and mouse Notch1 (ref. 22) reduce the ability of receptor activity to induce cell fate changes, as also found in *C. elegans*. Last, mouse Notch1 ANK mutants fail to induce target gene transcription^{23,24}. As the *C. elegans* ANK mutations abrogate ternary complex formation, we predict that ANK mutations in other Notch-related receptors similarly affect complex formation with a LAG-3-like protein and the CSL DNA-binding protein. That LAG-3 activity might be provided by a component of the Notch pathway, such as *mastermind*²⁵, which so far has not found a place in models for molecular mechanism, or some other protein that has yet to be identified. Identification of the third partner and analysis of the ternary complex will be essential for understanding how normal Notch signalling controls development and physiology and how aberrant Notch signalling leads to human disease.

Note added in proof: *lag-3* has also been shown to correspond to the previously described locus *sel-8* and to be associated with *Lag* phenotypes in RNAi experiments (T. Doyle and I. Greenwald, personal communication). □

Methods

Yeast two-hybrid screen and assays

The yeast two-hybrid screen was done essentially as described²⁶ with two modifications. First, we used the L40-*ura3* (*MATa, trp1, his3, leu2, ade2, LYS2::LexAop4-HIS3, ura3::LexAop8-lacZ*) yeast strain (a gift from T. Triolo and R. Sternberg) for both screening and characterization of interactions. Second, in addition to LexA and AD fusion proteins, a third, 'non-hybrid' protein, LAG-1, was expressed from pYES2 vector (Invitrogen). The screen used the plasmid form of a *C. elegans* cDNA library fused to Gal4-AD, which was derived from the LambdaACT-PRB-2 library²⁷. Among the six *lag-3* cDNAs recovered, three were full-length for LAG-3A (residues 1–490) and the others encoded truncated forms (residues 1–398, 1–446 and 24–490). The β -galactosidase filter-lift assay was done as described²⁶. For assays of yeast grown at different temperatures, yeast were grown on plates at the given temperature, frozen in liquid nitrogen and the enzymatic assay carried out at 30 °C.

Plasmids

Plasmids were constructed using standard techniques and details are available upon request. The pBTM116 vector²⁶ was used to obtain LexA fusions. All LexA-GLP-1 (RAM-ANK) constructs encode GLP-1 residues 788–1171; mutant versions contain appropriate nucleotide substitutions¹³. LexA-LIN-12 (RAM-ANK) constructs encode LIN-12 residues 930–1320, the mutant version contains the appropriate nucleotide substitution¹². LexA-LAG-3 encodes residues 1–490 of LAG-3. The LexA-Lamin fusion has been described²⁶. The non-hybrid used for the yeast two-hybrid screen contained residues 203–673 of LAG-1. AD-LAG-3 contains residues 1–490 of LAG-3A. pCITE-4a (Novagen) was used to obtain S-tag fusion proteins. S-LAG-1 includes residues 199–673 of LAG-1. S-GLP-1 (RAM-ANK), wild type and mutant, encodes residues 788–1171 of GLP-1. S-LAG-3 encodes residues 5–490 of LAG-3A. pET-28a (Novagen) was used to generate T7-tag fusions. T7-LAG-1 encodes residues 199–673 of LAG-1; T7-LAG-3 encodes residues 5–490 of LAG-3. For LAG-3 localization in worms, a Myc-LAG-3 construct, encoding full-length LAG-3A, was created from a modified pPD49.83, which contains the heat shock promoter HSP16-14, a translation start site, and an amino-terminal c-Myc epitope⁸.

In vitro binding assay

T7-tagged proteins were produced in *Escherichia coli* and partially purified by absorption onto T7-tag antibody agarose. S-tagged proteins were generated using rabbit reticulocyte lysate (TnT, Promega). *In vitro* translated proteins were diluted with 0.5 ml of T7-tag bind/wash buffer (4.29 mM Na₂HPO₄, 1.47 mM KH₂PO₄, 2.7 mM KCl, 137 mM NaCl, 0.1% Tween-20, 0.002% NaN₃, pH 7.3) in the presence of 1 × 'Complete EDTA-free' protease inhibitor cocktail (Boehringer Mannheim). S-tagged proteins were then added to the corresponding T7-tagged protein bound to the agarose beads in microfuge tubes. Samples were incubated with shaking for 1 h at 25 °C. The agarose beads were washed three times with 1 ml of T7-tag bind/wash buffer in the presence of 1 × Complete EDTA-free protease inhibitor cocktail. Bound proteins were released by boiling agarose beads in protein loading buffer. Proteins were separated using SDS-PAGE and transferred to polyvinylidene difluoride membrane. S-tagged proteins were visualized using S-protein-alkaline phosphatase conjugates (Novagen); T7-tagged proteins were visualized using T7-tag antibody-alkaline phosphatase conjugates (Novagen). Between 0.5 µg and 5 µg of T7-LAG-1 or T7-LAG-3 and about 100 ng of each of the S-tagged proteins were used in each assay.

Transcript analysis

The 5' end of the *lag-3* message was determined by RT-PCR using *lag-3* specific primers and a primer to the *trans*-splicing leader SL1. Two alternative 5' ends were obtained. The inferred transcripts are predicted to encode two proteins, LAG-3A and LAG-3B.

RNA interference

The RNAi effect should be limited to *lag-3* specifically, because this gene is unique in the *C. elegans* genome²⁸. The DNA template for *in vitro* transcription contained all of the coding sequence for LAG-3A and 86 nucleotides of the *lag-3* 3' untranslated region. To assay the postembryonic *lag-3* phenotype, L1 larvae were soaked in 1–3 µg µl⁻¹ of double stranded *lag-3* RNA for 48 h at 20 °C or 25 °C in M9 buffer (22 mM KH₂PO₄, 42 mM NaH₂PO₄, 85 mM NaCl, 1 mM MgSO₄) in the presence of *E. coli* (Absorbance 600 nm, 0.75–1.0). After soaking, animals were transferred to Petri dishes to continue development. To assay the embryonic *lag-3* phenotype, L4 animals were soaked in M9 containing 1 µg µl⁻¹ *lag-3* dsRNA overnight, then transferred to plates to lay eggs, using essentially a described method²⁹.

Subcellular localization

Generation of transgenic nematodes carrying HS-Myc-LAG-3, heat shock conditions and antibody detection was done essentially as described⁸, except a whole mount freeze cracking method was used for fixation of the worms³⁰.

Received 25 January; accepted 3 March 2000.

- Kimble, J. & Simpson, P. The LIN-12/Notch signalling pathway and its regulation. *Annu. Rev. Cell Dev. Biol.* **13**, 333–361 (1997).
- Artavanis-Tsakonas, S., Rand, M. D. & Lake, R. J. Notch signalling: cell fate control and signal integration in development. *Science* **284**, 770–776 (1999).
- Ellisen, L. W. *et al.* TAN-1, the human homolog of the *Drosophila* Notch gene, is broken by chromosomal translocations in T lymphoblastic neoplasms. *Cell* **66**, 649–661 (1991).
- Joutel, A. *et al.* Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia. *Nature* **383**, 707–710 (1996).
- Schroeter, E. H., Kisslinger, J. A. & Kopan, R. Notch1 signalling requires ligand-induced proteolytic release of intracellular domain. *Nature* **393**, 382–386 (1998).
- Struhl, G. & Adachi, A. Nuclear access and action of Notch *in vivo*. *Cell* **93**, 649–660 (1998).
- Lecourtis, M. & Schweisguth, F. Indirect evidence for Delta-dependent intracellular processing of notch in *Drosophila* embryos. *Curr. Biol.* **8**, 771–774 (1998).
- Roehl, H., Bosenberg, M., Blöchl, R. & Kimble, J. Roles of the RAM and ANK domains in signalling by the *C. elegans* GLP-1 receptor. *EMBO J.* **15**, 7002–7012 (1996).
- Tamura, K. *et al.* Physical interaction between a novel domain of receptor Notch and the transcription factor RBP-Jk/Su(H). *Curr. Biol.* **5**, 1416–1423 (1995).
- Christensen, S., Kodoyianni, V., Bosenberg, M., Friedman, L. & Kimble, J. *lag-1*, a gene required for *lin-12* and *glp-1* signalling in *Caenorhabditis elegans*, is homologous to human CBF1 and *Drosophila* Su(H). *Development* **122**, 1373–1383 (1996).
- Roehl, H. & Kimble, J. Control of cell fate in *C. elegans* by a GLP-1 peptide consisting primarily of ankyrin repeats. *Nature* **364**, 632–635 (1993).
- Greenwald, I. & Seydoux, G. Analysis of gain-of-function mutations of the *lin-12* gene of *Caenorhabditis elegans*. *Nature* **346**, 197–199 (1990).
- Kodoyianni, V., Maine, E. M. & Kimble, J. Molecular basis of loss-of-function mutations in the *glp-1* gene of *Caenorhabditis elegans*. *Mol. Biol. Cell* **3**, 1199–1213 (1992).
- Mango, S. E., Maine, E. M. & Kimble, J. Carboxy-terminal truncation activates *glp-1* protein to specify vulval fates in *Caenorhabditis elegans*. *Nature* **352**, 811–815 (1991).
- Austin, J. & Kimble, J. *glp-1* is required in the germ line for regulation of the decision between mitosis and meiosis in *C. elegans*. *Cell* **51**, 589–599 (1987).
- Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998).
- Greenwald, I. S., Sternberg, P. W. & Horvitz, H. R. The *lin-12* locus specifies cell fates in *Caenorhabditis elegans*. *Cell* **34**, 435–444 (1983).
- Berry, L. W., Westlund, B. & Schedl, T. Germ-line tumor formation caused by activation of *glp-1*, a *Caenorhabditis elegans* member of the Notch family of receptors. *Development* **124**, 925–936 (1997).
- Mitchell, P. J. & Tjian, R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245**, 371–378 (1989).
- Hicks, G. R. & Raikhel, N. Y. Protein import into the nucleus: an integrated view. *Annu. Rev. Cell Dev. Biol.* **11**, 155–188 (1995).

- Diederich, R. J., Matsuno, K., Hing, H. & Artavanis-Tsakonas, S. Cytosolic interaction between deltex and Notch ankyrin repeats implicates deltex in the Notch signalling pathway. *Development* **120**, 473–481 (1994).
- Kopan, R., Nye, J. S. & Weintraub, H. The intracellular domain of mouse Notch: a constitutively activated repressor of myogenesis directed at the basic helix-loop-helix region of MyoD. *Development* **120**, 2385–2396 (1994).
- Jarriault, S. *et al.* Signalling downstream of activated mammalian Notch. *Nature* **377**, 355–358 (1995).
- Kato, H. *et al.* Involvement of RBP-J in biological functions of mouse Notch1 and its derivatives. *Development* **124**, 4133–4141 (1997).
- Smoller, D. *et al.* The *Drosophila* neurogenic locus *mastermind* encodes a nuclear protein unusually rich in amino acid homopolymers. *Genes Dev.* **4**, 1688–1700 (1990).
- Bartel, P. L. & Fields, S. (eds) *The Yeast Two-Hybrid System* (Oxford Univ. Press, New York, 1997).
- Kraemer, B. *et al.* NANOS-3 and FBF proteins physically interact to control the sperm-oocyte switch in *Caenorhabditis elegans*. *Curr. Biol.* **9**, 1009–1018 (1999).
- The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998); erratum *ibid* **283**, 35 (1999); erratum *ibid* **283**, 2103 (1998).
- Subramaniam, K. & Seydoux, G. *nos-1* and *nos-2*, two genes related to *Drosophila* nanos, regulate primordial germ cell development and survival in *Caenorhabditis elegans*. *Development* **126**, 4861–4871 (1999).
- Crittenden, S. L. & Kimble, J. in *Cell: A Laboratory Manual* (eds Spector, D., Goldman, R. & Lewin, L.) 108.101–108.109 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1998).

Acknowledgements

We acknowledge A. Puoti for providing the plasmid form of the cDNA library; R. Sternglanz, S. M. Hollenberg and A. Grimson for yeast strains and plasmids; and A. Steinberg and L. Vanderploeg for help with the illustrations. A.G.P. is an Howard Hughes Medical Institute predoctoral fellow. J.K. is an investigator of the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to J.K. (e-mail: jekimble@facstaff.wisc.edu). The mRNA sequences for *lag-3A* and *lag-3B* have been deposited in GenBank under accession codes AF241846 and AF241847, respectively.

Translocation step size and mechanism of the RecBC DNA helicase

Piero R. Bianco & Stephen C. Kowalczykowski

Sections of Microbiology and of Molecular and Cellular Biology, University of California at Davis, Davis, California 95616, USA

DNA helicases are ubiquitous enzymes that unwind double-stranded DNA^{1–3}. They are a diverse group of proteins that move in a linear fashion along a one-dimensional polymer lattice—DNA—by using a mechanism that couples nucleoside triphosphate hydrolysis to both translocation and double-stranded DNA unwinding to produce separate strands of DNA. The RecBC enzyme is a processive DNA helicase that functions in homologous recombination in *Escherichia coli*; it unwinds up to 6,250 base pairs per binding event and hydrolyses slightly more than one ATP molecule per base pair unwound. Here we show, by using a series of gapped oligonucleotide substrates, that this enzyme translocates along only one strand of duplex DNA in the 3' → 5' direction. The translocating enzyme will traverse, or 'step' across, single-stranded DNA gaps in defined steps that are 23 (±2) nucleotides in length. This step is much larger than the amount of double-stranded DNA that can be unwound using the free energy derived from hydrolysis of one molecule of ATP, implying that translocation and DNA unwinding are separate events. We propose that the RecBC enzyme both translocates and unwinds by a quantized, two-step, inchworm-like mechanism that may have parallels for translocation by other linear motor proteins.

DNA unwinding by RecBC enzyme initiates only at blunt or nearly blunt double-stranded DNA (dsDNA) ends; thus, it had not been possible to determine on which strand of DNA translocation occurs, or whether both DNA strands are required. We therefore

devised a strategy to determine both the polarity of and the step size for translocation. We first introduced a series of defined-size gaps in one or the other strand of otherwise duplex DNA, and then determined which size single-stranded DNA (ssDNA) gap prevented traversal. We used two types of substrates, each with a double-stranded blunt-end entry site at the 'proximal' 25-mer, an ssDNA gap, a 'tester' oligonucleotide (the 20-mer) distal to the gap and an ssDNA tail downstream of the annealed 20-mer (Fig. 1). The tail limits entry to only the dsDNA end; the gap is used to define the importance of that DNA strand in helicase translocation; and the displacement of the distal, tester oligonucleotide reports translocation and unwinding past the gap. Substrate set A has gaps positioned in the 5'-terminated strand relative to the entry point of the enzyme (the bottom strand of the substrates as shown), and substrate set B has gaps positioned in the 3'-terminated strand relative to the entry point of the enzyme (the top strand as shown). Gaps were initially either 0 (a nick) or 30 nucleotides (nt) in size (Fig. 1). A gap size of 30 nt was selected, as this is 9 nt larger than the contact length of the stationary holoenzyme bound to a dsDNA end⁴.

We reasoned that the translocating enzyme should dissociate from the substrate when it encounters a discontinuity larger than its contact length in a strand of DNA that is critical to translocation. If the enzyme translocates along only one strand of DNA, then it

should fail to bypass a gap present in that strand but not the other; if the enzyme requires both strands of DNA for translocation, then it should fail to bypass a gap (larger than its contact length) present in either strand. Failure to bypass a gap would be observed as formation of an intermediate unwinding product, consisting of the 100-mer (or alternatively the 130-mer) with the tester 20-mer distal to the gap, still bound.

When gaps of either 0 or 30 nt are present in the bottom strand, RecBC enzyme unwinds the substrate, displacing both the 25-mer proximal to the gap and the 20-mer downstream (Fig. 1a). This is observed as the disappearance of substrate in a time-dependent manner concomitant with the production of 100-mer ssDNA. Furthermore, displacement of both oligonucleotides is simultaneous (data not shown) with little or no intermediate species produced for either of these substrates, showing that a second RecBC enzyme is not responsible for displacement of the distal oligonucleotide. In the presence of single-stranded DNA binding (SSB) protein, RecBC enzyme cannot unwind substrates with only the internal 20-mers bound to the 100-mer (that is, substrates with a ssDNA tail at each end), showing that the enzyme requires the proximal dsDNA region to mediate unwinding beyond the gap (data not shown). Thus, RecBC enzyme is able to bypass gaps present in the 5'-terminated, or bottom strand, with more than 92% efficiency.

When gaps are positioned in the 3'-terminated strand, however, the results are significantly different (Fig. 1b). For the 0-nt-gap substrate, the enzyme can unwind and displace both the proximal 25-mer and the distal 20-mer; no intermediate is produced and ~65% of the starting substrate is converted to 100-mer in 2 min. In contrast, for the 30-nt-gap substrate the enzyme can displace only the proximal 25-mer; the 20-mer distal to the gap is not displaced efficiently, a significant amount of intermediate species accumulates (~55%), and the amount of 100-mer produced is decreased (down from 65% to 16%). Thus, RecBC enzyme is unable efficiently to bypass a 30-nt gap when that gap is present in the 3'-terminated, or top strand.

Identical results were obtained using single-round kinetic experiments in which the enzyme was pre-bound to the dsDNA end, and the reactions were initiated with a mixture of ATP and heparin (data not shown). Heparin prevents re-binding of RecBC enzyme after it has dissociated from a DNA molecule (ref. 5; and data not shown). Thus, the ability of the enzyme to bypass the 30-nt gap in the bottom strand (versus the top strand) is not owing to dissociation of

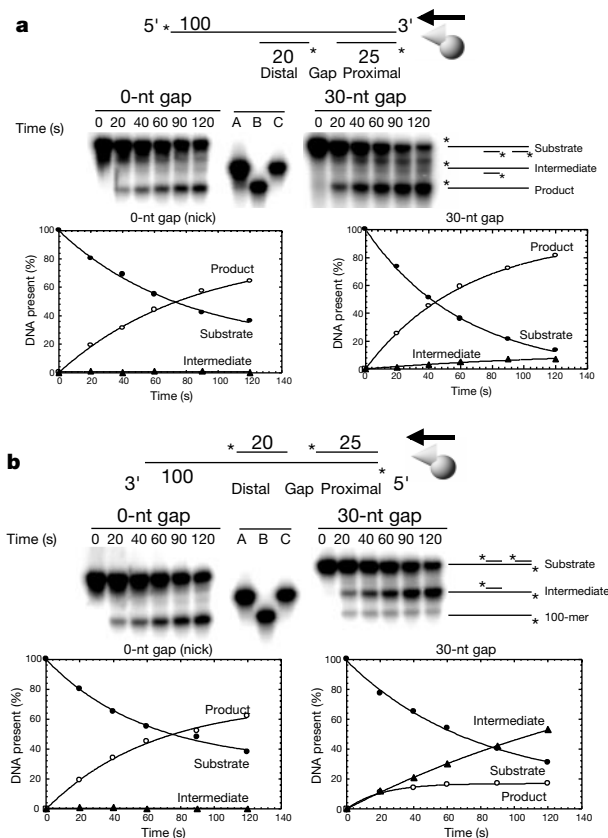


Figure 1 The RecBC helicase translocates in the 3' → 5' direction. Time courses of unwinding reactions using substrates with gaps present in the 5'-terminated (bottom) strand (a) and in the 3'-terminated (top) strand (b) are shown. The substrate for each reaction is shown above the gel. RecBC enzyme must translocate from right to left as shown (arrow); displacement of the oligonucleotide distal to the gap (20-mer) requires both displacement of the oligonucleotide proximal to the gap (25-mer) and traversal through the gap. The positions of substrate (filled circles), intermediate (triangles) and 100-mer product (open circles) are indicated to the right of each gel. Lanes A, B and C are standards: A, intermediate for the nicked substrate; B, 100-mer product; C, the intermediate for the 30-nt gap.

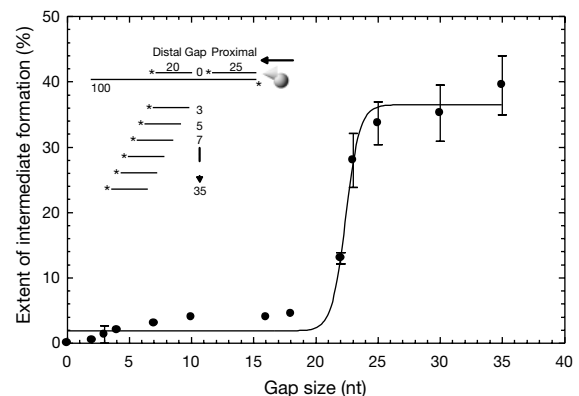


Figure 2 The observed step size for the translocating RecBC enzyme is 23 (±2) nucleotides. A compilation of the extents of intermediate formation from two-minute time courses for the substrates with increasing gap lengths is shown. Inset shows the substrates, which consist of a 100-mer with a 25-mer annealed to the 5' end, and with various 20-mers annealed to different positions on the 100-mer. Reactions and data analyses were carried out as in Fig. 1.

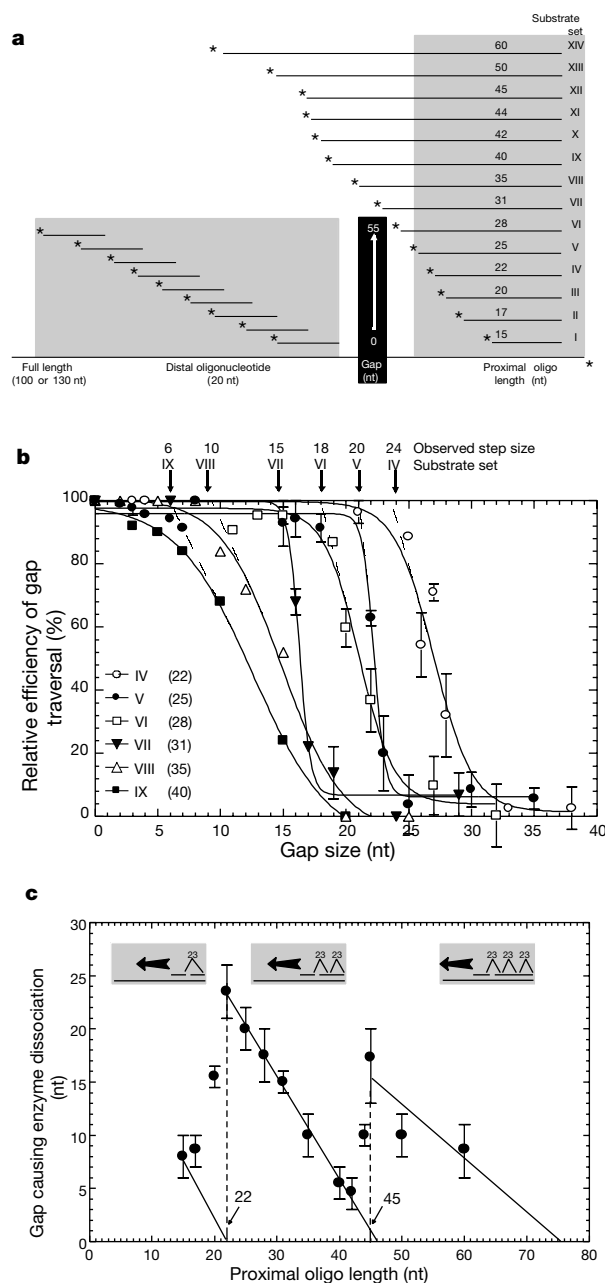


Figure 3 The observed step size is dependent on the length of the proximal duplex DNA region and displays a 23-nt periodicity for 3 cycles of translocation and DNA unwinding. **a**, Substrates. The arrangement of proximal and distal oligonucleotides bound to same 100-mer (or 130-mer) is shown. Thirteen additional substrate sets were constructed in identical fashion to those in Fig. 2 and are arranged according to the length of the oligonucleotide proximal to the gap (right grey box). In each set, the sequences of the 100-mer (or 130-mer, with 30 additional bases at the 3' end) and the set of 20-mers (left grey box) were always the same. The potential gap length generated is indicated (black box). **b**, Data compiled from helicase assays such as those shown in Fig. 2 using substrate sets IV–IX. Data are shown for a two-minute time point, averaged from two to seven assays, done on separate days for each substrate. For each substrate set, the extents were normalized to the average maximum extent of intermediate formation for that set: set IV, 56%; set V, 40%; set VI, 55%; set VII, 40%; and sets VIII and IX, 50%. Curves represent a fit to a sigmoid function (Prism v3.0, GraphPad Software) and were used to obtain the observed step size. Dashed lines and arrows indicate the observed step size for each data set (Table 1). Numbers in parentheses indicate the length of the proximal oligonucleotide. **c**, Compilation of data from helicase assays using substrate sets I–XIV. The gap causing enzyme dissociation for each substrate set (proximal oligonucleotide length) is the gap length at which the curves shown in **b** begin to decrease. Data for substrate sets I–III and X–XIV are not shown. Error bars represent the minimum and maximum value of the extrapolated step size. Grey boxes indicate the number of translocation steps taken by the helicase.

the enzyme at the gap followed by rapid re-binding with continued translocation. Furthermore, the inability to bypass gaps is not owing to a failure of RecBC enzyme to displace SSB protein bound in the gap, as similar results were obtained using the bacteriophage T4 gene 32 protein (data not shown). Collectively, these data show that, although it initiates by binding to dsDNA, RecBC enzyme requires only one strand of dsDNA for translocation (the strand 3' terminated at the entry end). Thus, RecBC enzyme translocates on this strand with 3' → 5' polarity.

Because the enzyme dissociates from the substrate when the strand on which it translocates has a gap of 30 nt, but not when the gap is a nick, this suggested a strategy to define the translocation step size of the enzyme. We reasoned that by varying the length of the gap in the 3'-terminated strand, we could determine the smallest gap size that the enzyme fails to bypass during translocation, and that this should correspond to the 'step size' for the enzyme. We therefore constructed a family of substrates with ssDNA gaps ranging from a nick to 35 nt (Fig. 2, inset).

Reactions and analyses identical to those in Fig. 1 were carried out for the substrates in Fig. 2. RecBC enzyme can bypass gaps up to 18 nt in size as determined by the low extents of intermediate formation. Above this gap size, the accumulation of intermediate increased ninefold to an average of 36% for gaps of 25 nt or bigger. Some RecBC enzymes may be able to traverse these larger gaps because of the inherent flexibility of the ssDNA regions in these gapped substrates; that is, the ssDNA in these gaps may 'loop out' bringing the two flanking dsDNA regions of the substrate into close proximity, thereby facilitating enzyme bypass. In any case, the sharp transition at 20–25 nt shows that translocation through such gaps is seriously impaired. Thus, these data show that when RecBC enzyme encounters gaps smaller than 22–23 nt, it can 'step across' them as though they were dsDNA, suggesting that the step size for the translocating enzyme is ~23 nt. In addition, these results imply the disposition of the DNA unwinding domain relative to the footprint of the enzyme: this domain must be positioned at or near the rear of the helicase for gap bypass to occur. If it were positioned at the front, then a gap in the translocating strand as small as 1 nt would function as a suicide substrate by causing the helicase to dissociate as a complex with the ssDNA that it had produced and released by DNA unwinding.

Such a large step size was unanticipated, as the largest unwinding step size reported to our knowledge for a DNA helicase is 4–5 nt (ref. 6). Furthermore, this translocation step size is much larger than the amount of DNA that can be unwound from the free energy derived from the hydrolysis of one molecule of ATP^{2,7}. To ensure that the failure of RecBC enzyme efficiently to bypass gaps of 23 (±2) nt or greater was not a consequence of the oligonucleotides used to construct the gap (because, for example, the sequence of the 20-mers following the gap is always different), additional substrate sets were constructed. For each set, the same 100-mer and family of distal 20-mers were used as for the substrates in Fig. 2, but the length of the oligonucleotide proximal to the gap was modified (Fig. 3a, substrate sets IV–IX). Thus, a gap of the same size would be present in each substrate set but the sequence of the oligonucleotides flanking that gap would be different. We expected that gap bypass should be the same in each of the six substrate sets, that is, ~23 nt.

Reactions and analyses were carried out using sets IV–IX; these data were converted to a 'relative efficiency of gap traversal' (Fig. 3b). Rather than being identical, each substrate set displays a distinctive curve. Unexpectedly, the gap-bypass efficiency decreases rapidly at a gap size that changes monotonically with each substrate set. Thus, the observed step size for the translocating RecBC enzyme appears to be dependent on the length of the proximal oligonucleotide (Table 1). This unexpected finding shows that the ability of RecBC helicase to traverse gaps as large as ~23 nt cannot be explained by proposing that the enzyme, with a static footprint of ~22 nt, simply positions itself across the gap so that the leading end

Table 1 Effects of proximal oligonucleotide length on gap bypass

Step cycle*	Substrate set	Proximal oligo length (nt)†	Observed step size (nt)‡	Total (nt)§	Inferred step size (nt) (mean = 23)
1	I	15	8	23	23
1	II	17	9	26	26
2	III	20	16	36	18
2	IV	22	24	46	23
2	V	25	20	45	23
2	VI	28	18	46	23
2	VII	31	15	46	23
2	VIII	35	10	45	23
2	IX	40	6	46	23
2	X	42	5	47	23
2	XI	44	10	54	27
3	XII	45	17	62	21
3	XIII	50	10	60	20
3	XIV	60	9	69	23

* Translocation step as shown in Fig. 3c.

† The length of the oligonucleotide proximal to the gap.

‡ The observed step size corresponds to the gap at which the relative efficiency of gap traversal decreases below 100%.

§ The total is the sum of the length of the oligonucleotide proximal to the gap and the observed step size.

|| The inferred step size is equal to the total for sets I and II, and is obtained by dividing the total by 2 for sets III–XI and by 3 for sets XII–XIV (see text).

can bind the dsDNA across any ~ 23 -nt gap. For such a model, in which the translocation step size is fixed relative to the end of the proximal dsDNA, RecBC enzyme should traverse any 23-nt gap, regardless of proximal dsDNA length. Thus, the data shown in Figs 2 and 3b argue against a model in which a motor domain at the rear of a fixed step-size helicase propels the leading edge of the protein across the gap where the leading domain can engage the duplex DNA, and thereby effect gap traversal. If this model were operative, then the gap size traversed should be a constant property (that is, 23 nt for all substrate sets) of the helicase.

The data in Table 1 (IV–IX) reveal that there is an inverse relationship between the size of the smallest gap that the enzyme fails to traverse, and the length of the oligonucleotide proximal to the gap. When these gap sizes are added to the length of the oligonucleotide proximal to the gap, a total length of 46 nt is obtained for each substrate set (column 4). This suggests that the step size is quantized and its length is determined relative to the entry point of RecBC enzyme into the dsDNA. In addition, the 3' end of the distal 20-mer is always 46 nt from the entry point when the blockage of translocation begins to occur. This implies either that the enzyme translocates with a step size of 46 nt, or that the step size is 23 nt and the enzyme fails to bypass the gap and dissociation occurs during the second step. The first interpretation is unlikely because the footprint of the holoenzyme enzyme is only 19 (± 3) base pairs (bp) (ref. 4) and it functions as a single multimeric complex containing only one ATPase subunit^{8–10}. Furthermore, if the step size were 46 nt, then for all substrates where the sum of the proximal and distal oligonucleotides and the gap was less, neither of the oligonucleotides would be displaced, because in the first step the enzyme would step into a gap beyond the bound oligonucleotides and dissociate. Clearly, this is not the case. The second interpretation is more likely, which argues that the step size for RecBC enzyme is 23 nt (Table 1, IV–IX, right column).

To determine whether the step size remains constant during translocation, we constructed two additional groups of substrates. The first group used the shortest proximal oligonucleotides that could be reliably bound to the 100-mer: these were 15, 17 and 20 nt (I–III, Fig. 3a). The second group used longer proximal oligonucleotides (42, 44, 45, 50 or 60 nt; X–XIV) and required a 130-mer oligonucleotide instead of the 100-mer for substrate sets XII–XIV. We reasoned that if the step size remained constant throughout translocation, then the observed gap size traversed should display periodicity. It should approach a minimum at a proximal oligonucleotide length of ~ 23 nt, increase abruptly to ~ 23 and then decrease linearly over the span of ~ 23 nt, reaching a second

minimum at a proximal oligonucleotide length of ~ 46 nt, and so on. The results of helicase assays using these substrates clearly show a cyclic pattern (Fig. 3c), in which the gap-size traversed approaches a minimum, increases to 24 nt, decreases to a new extrapolated minimum at 46 nt, and then is followed by yet another sharp increase and another (partial) linear decline; that is, the cycle resets approximately every 23 nt. Thus, the step size of ~ 23 nt is constant for at least three cycles of translocation.

Our finding that RecBC enzyme can traverse ssDNA gaps smaller than 23 nt in length requires a mechanism to explain how a helicase can translocate with a step size of 23 despite being limited, thermodynamically, to no more than ~ 5 bp of unwinding per mole of ATP hydrolysed. To explain this potential dilemma, we propose a mechanism that we have termed the 'quantum inchworm' (Fig. 4), which is similar to a model proposed for *E. coli* RNA polymerase¹¹. In the quantum inchworm model, translocation and DNA unwinding are two separate, consecutive events: first, translocation occurs by a 23-bp 'step'; second, unwinding at the trailing domain occurs by several, smaller events of 2–5 bp for each ATP molecule hydrolysed. Consistent with the latter proposal, kinetic measurement of the DNA unwinding step size for RecBCD enzyme shows that it unwinds dsDNA in increments of 4–5 bp (A. Vindigni, personal communication). We propose that RecBC enzyme possesses two, non-equivalent DNA-binding sites: the first DNA-binding site is the leading domain of the enzyme that binds to one strand (the 3'-terminated strand) of dsDNA and functions to anchor the enzyme during DNA unwinding. The second DNA-binding site is the trailing domain that functions as the helicase domain of the enzyme and is responsible for separating the strands of DNA; this second domain must move in several smaller steps relative to the leading domain. We propose that one complete cycle of translocation and DNA unwinding is achieved by the expansion and contraction of the enzyme that includes the hydrolysis of at least 5 to 12 ATP molecules⁷.

The cycle (Fig. 4) begins with the enzyme bound to a dsDNA end (stage 1). The trailing domain is bound at the end and, consequently, the leading edge is positioned 23 nt from this end (23 bp in B-form dsDNA corresponds to ~ 7.8 nm). The leading domain anchors the enzyme in place by binding to only one strand of the DNA duplex (indicated by the hands wrapping around one strand of the duplex). The trailing domain unwinds the dsDNA behind the

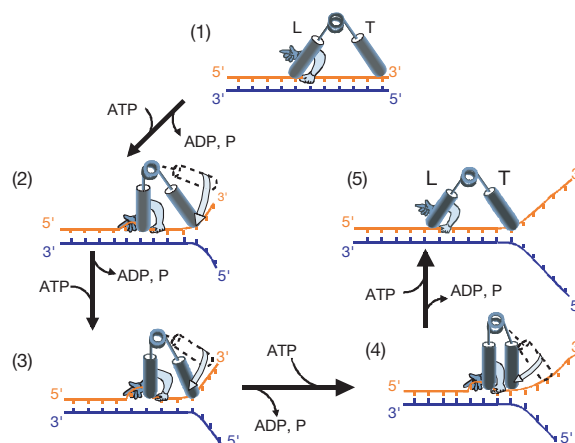


Figure 4 The RecBC enzyme translocates by a 'quantum inchworm' mechanism. A single catalytic cycle of translocation and DNA unwinding is shown. The individual strands of the DNA duplex are coloured orange (3'-terminated strand) and blue (5'-terminated strand). Translocation is from right to left along only one strand (3'-terminated) of the duplex. The enzyme contains two domains: a leading domain (L, which anchors the enzyme to only one strand of duplex DNA) and a trailing domain (T, responsible for DNA unwinding). The leading domain binds 23 nt ahead of the trailing domain.

leading edge, using the energy derived from ATP binding/hydrolysis both to separate the DNA strands and to advance with 4–5-bp steps toward the leading domain (stages 2–4). As the lagging domain translocates up to the leading domain (stage 4), a signal (yet to be determined) is transmitted from the lagging to the leading domain causing the leading domain to dissociate and rebind ~23 nt ahead of the bound lagging end (stage 5). The enzyme has now returned to the starting point and the cycle repeats itself until the DNA molecule is completely unwound or the enzyme dissociates. As the lagging domain approaches the anchored leading domain, torsional stress will accumulate in the intervening dsDNA. Because the trailing end of the enzyme must transiently disengage to advance, any accumulated torsional stress in the DNA could be released at this point, and, provided that the trailing end rapidly engages before the strands fully re-anneal, helicase function will remain processive as long as the leading end remains attached. One might expect to see an occasional uncoupling of unwinding and translocation (for example, the DNA strands partially re-anneal before the trailing end engages; this would result in an ‘inefficiency’ in ATP utilization. In fact, such an inefficiency is observed: for RecBC enzyme, as many as 1.4 ATP molecules are hydrolysed per base pair unwound¹², whereas for RecBCD enzyme 2–3 ATP molecules are hydrolysed per base pair unwound⁷; this is between 3–15-fold more hydrolysis than the minimum required of 0.2 ATP molecules hydrolysed per base pair unwound. Therefore, at most, only 1 of 3 hydrolytic events results in DNA unwinding, and the value may be as few as 1 of 15 events being productive. Thus, anchoring of the helicase to DNA by the leading domain gives the trailing unwinding domain multiple opportunities to act on its substrate without macroscopic dissociation.

The large conformational changes proposed are not without precedent. The crystal structures of the *Bacillus stearothermophilus* PcrA helicase^{13,14} and the *E. coli* Rep helicase¹⁵ have been determined. PcrA helicase was proposed to function as an inchworm¹⁶, as was the related UvrD helicase¹⁷. Our mechanism has parallels to those proposed for these helicases and, thus, we conclude that a similar underlying mechanism may be used by at least a subset of the linear motor proteins to perform mechanical work. □

Methods

Substrate construction

Oligonucleotides used for substrate construction were purified using denaturing polyacrylamide gels. The sequence of the 130-mer is 5′-TGGCTGCAACGCGGCATCCCCGATGCGCCGGAAGCGAGAAGAATCATAATGGGGAAGGCCACCGCTCGCGT CGCGAACGCCAGCAAGACGTAGCCAGCGCGTGGCCGCCATGCCGCGCAT AATG-3′; the 100-mer lacked the last 30 nt. Oligonucleotide concentrations were determined using an extinction coefficient calculated for each oligonucleotide. Oligonucleotides were labelled at the 5′ ends using T4 polynucleotide kinase and γ -³²P-ATP. Annealing was performed in T4 polynucleotide kinase buffer plus 10 mM magnesium acetate containing the 100- (or 130-), 25- and 20-mers at a ratio of 100-mer:25-mer:20-mer of 1:1.1:1.2, to ensure that all of the 100-mer was converted to the substrate form. The mixture was heated to 100 °C for 5 min, and allowed to cool to room temperature. Substrate formation was confirmed by electrophoresis. Typically, 100% of the 100-mer was converted to substrate; the SSB protein bound the residual 25- and 20-mer.

Assay system

Helicase assays were conducted at room temperature and contained 25 mM Tris-HCl pH 7.5, 1 mM dithiothreitol, 8 mM magnesium acetate, 1 mM ATP, 1 μ M SSB, 10 nM DNA (in mol of 100-mers) and 1.03 nM active RecBC enzyme. SSB protein ensures entry from the blunt end of the DNA, and prevents re-annealing of unwound strands. For single-round experiments, RecBC helicase was bound to DNA for 2 min and reactions were initiated by the addition of an ATP-heparin mix (1 mM and 10 mg ml⁻¹, final). Reactions were stopped by an equal volume of ficoll gel loading dye mix, containing SDS (1%, final), EDTA (50 mM) and proteinase K (0.4 mg ml⁻¹). After 5 min, aliquots were loaded onto polyacrylamide gels (1:1.5 bis:acrylamide in TBE buffer). After electrophoresis, the gels were analysed using a Molecular Dynamics Storm 840 and ImageQuaNT software.

Received 4 October; accepted 15 March 2000.

1. Matson, S. W., Bean, D. W. & George, J. W. DNA helicases: enzymes with essential roles in all aspects of DNA metabolism. *Bioessays* **16**, 13–22 (1994).
2. Lohman, T. M. & Bjornson, K. P. Mechanisms of helicase-catalyzed DNA unwinding. *Annu. Rev. Biochem.* **65**, 169–214 (1996).

3. West, S. C. DNA helicases: new breeds of translocating motors and molecular pumps. *Cell* **86**, 177–80 (1996).
4. Ganesan, S. & Smith, G. R. Strand-specific binding to duplex DNA ends by the subunits of *Escherichia coli* RecBCD enzyme. *J. Mol. Biol.* **229**, 67–78 (1993).
5. Korangy, F. & Julin, D. A. Kinetics and processivity of ATP hydrolysis and DNA unwinding by the RecBC enzyme from *Escherichia coli*. *Biochemistry* **32**, 4873–4880 (1993).
6. Ali, J. A. & Lohman, T. M. Kinetic measurement of the step size of DNA unwinding by *Escherichia coli* UvrD helicase. *Science* **275**, 377–380 (1997); erratum, *ibid* **276**, 21 (1997).
7. Roman, L. J. & Kowalczykowski, S. C. Characterization of the adenosinetriphosphatase activity of the *Escherichia coli* RecBCD enzyme: Relationship of ATP hydrolysis to the unwinding of duplex DNA. *Biochemistry* **28**, 2873–2881 (1989).
8. Korangy, F. & Julin, D. A. A mutation in the consensus ATP-binding sequence of the RecD subunit reduces the processivity of the RecBCD enzyme from *Escherichia coli*. *J. Biol. Chem.* **267**, 3088–3095 (1992).
9. Taylor, A. F. & Smith, G. R. Monomeric RecBCD enzyme binds and unwinds DNA. *J. Biol. Chem.* **270**, 24451–24458 (1995).
10. Bianco, P. R. & Kowalczykowski, S. C. The recombination hotspot Chi is recognized by the translocating RecBCD enzyme as the single strand of DNA containing the sequence 5′-GCTGGTGG-3′. *Proc. Natl Acad. Sci. USA* **94**, 6706–6711 (1997).
11. Chamberlin, M. J. New models for the mechanism of transcription elongation and its regulation. *Harvey Lectures* **88**, 1–21 (1992).
12. Korangy, F. & Julin, D. A. Efficiency of ATP hydrolysis and DNA unwinding by the RecBC enzyme from *Escherichia coli*. *Biochemistry* **33**, 9552–9560 (1994).
13. Subramanya, H. S., Bird, L. E., Brannigan, J. A. & Wigley, D. B. Crystal structure of a DExx box DNA helicase. *Nature* **384**, 379–383 (1996).
14. Velankar, S. S., Soultanas, P., Dillingham, M. S., Subramanya, H. S. & Wigley, D. B. Crystal structures of complexes of PcrA DNA helicase with a DNA substrate indicate an inchworm mechanism. *Cell* **97**, 75–84 (1999).
15. Korolev, S., Hsieh, J., Gauss, G. H., Lohman, T. M. & Waksman, G. Major domain swiveling revealed by the crystal structures of complexes of *E. coli* Rep helicase bound to single-stranded DNA and ADP. *Cell* **90**, 635–647 (1997).
16. Bird, L. E., Subramanya, H. S. & Wigley, D. B. Helicases: a unifying structural theme? *Curr. Opin. Struct. Biol.* **8**, 14–18 (1998).
17. Mechanic, L. E., Hall, M. C. & Matson, S. W. *Escherichia coli* DNA helicase II is active as a monomer. *J. Biol. Chem.* **274**, 12488–12498 (1999).

Acknowledgements

We thank D. Anderson, R. Ando, D. Arnold, C. Barnes, R. Baskin, F. Chedin, F. Harmon, J. Kleiman, J. New, E. Seitz, and S. Shetterley for their comments. This work was supported by a grant from the NIH.

Correspondence and requests for materials should be addressed to S.C.K. (e-mail: sckowalczykowski@ucdavis.edu).

The crystal structure of the photoprotein aequorin at 2.3 Å resolution

James F. Head*, Satoshi Inouye†, Katsunori Teranishi‡ & Osamu Shimomura*§

* Structural Biology Group, Department of Physiology, Boston University School of Medicine, Boston, Massachusetts 02118, USA

† Yokohama Research Centre, Chisso Corporation, 5-1 Okawa, Kanazawa-ku, Yokohama 236, Japan

‡ Faculty of Bioresources, Mie University, Kamihama, Tsu 514, Japan

§ Marine Biologicals Laboratory, Woods Hole, Massachusetts 02543, USA

Aequorin is a calcium-sensitive photoprotein originally obtained from the jellyfish *Aequorea aequorea*¹. Because it has a high sensitivity to calcium ions and is biologically harmless, aequorin is widely used as a probe to monitor intracellular levels of free calcium. The aequorin molecule contains four helix–loop–helix ‘EF-hand’ domains, of which three can bind calcium². The molecule also contains coelenterazine as its chromophoric ligand³. When calcium is added, the protein complex decomposes into apoaequorin, coelenteramide and CO₂, accompanied by the emission of light⁴. Apoaequorin can be regenerated into active aequorin in the absence of calcium by incubation with coelenterazine, oxygen and a thiol agent⁵. Cloning and expression of the com-

plementary DNA for aequorin were first reported in 1985 (refs 2, 6), and growth of crystals of the recombinant protein has been described⁷; however, techniques have only recently been developed to prepare recombinant aequorin of the highest purity⁸, permitting a full crystallographic study. Here we report the structure of recombinant aequorin determined by X-ray crystallography. Aequorin is found to be a globular molecule containing a hydrophobic core cavity that accommodates the ligand coelenterazine-2-hydroperoxide. The structure shows protein components stabilizing the peroxide and suggests a mechanism by which calcium activation may occur.

Although aequorin is believed to be a functional monomer, there are two molecules A and B in the crystal structure. Figures shown here represent molecule A, but there are few differences between the molecules. Some surface side chains show different orientations, and part of an empty calcium-binding loop, residues 153–157, also appears to be in a different conformation. In molecule B, this loop is well defined and stabilized by crystal packing contacts, whereas in molecule A parts of the solvent-exposed loop are disordered and there is an apparent shift, by up to 10 Å, in the main-chain position. The differences in the conformation of this surface loop appear to have little effect on other parts of the molecule. Excluding residues 153–157, superimposition of molecules A and B gives r.m.s. deviations of 0.3 Å for main-chain atoms and 0.6 Å for all atoms. The recombinant protein used in this study also includes an amino-terminal modification, ANSKL-protein instead of VKL-protein, in the natural gene product; the remainder of the protein is identical². To avoid confusion with previous numbering in the literature, we retain the sequence numbering of the natural gene product. The N-terminal residues ANSK appear disordered in both molecules in the present structure. The residues of the molecules described are therefore numbered from 3 to 189.

The protein scaffold of aequorin consists of four EF-hand domains arranged in pairs to form the globular molecule. Each pair of 'hands' is arranged back-to-back, forming short stretches of β -sheet, as seen in other EF-hand proteins. Although it is in the calcium-free state, the whole globular aequorin molecule is formed in a manner rather analogous to calmodulin binding to its targets in the presence of calcium. A hydrophobic surface is present in each half of the molecule, situated within a cup formed by each pair of EF hands. As in calmodulin, these two cups are brought towards one another by the bend of a loop separating the second helix of EF hand II from the first helix of EF hand III (Fig. 1). Unlike calmodulin, the two cups of aequorin come together in a mouth-to-mouth alignment, rather than in the off-set arrangement of calmodulin gripping its targets. The loop connecting the two halves of the molecule also shows no evidence of the extreme flexibility of the linker of calmodulin. The two facing cups in aequorin form the coelenterazine-binding cavity. This arrangement of an EF-hand protein with an enclosed hydrophobic binding site is somewhat analogous to that of recoverin, in which a myristoyl 'tail' is tucked inside the protein in the calcium-free state⁹. In recoverin, however, the EF hands come together in a rather different organization in which only three of the

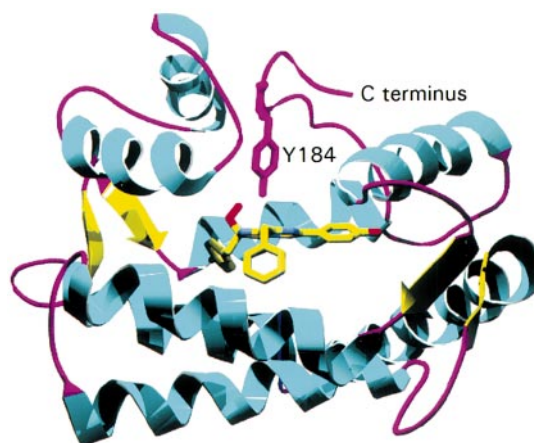


Figure 1 Ribbon representation showing the secondary structure elements in the protein. α -Helices are denoted in cyan, β -sheets in yellow, loops in magenta. Coelenterazine and the side chain of Tyr 184 are shown as stick representations. Figure prepared using Swiss-PdbViewer²⁹ and rendered using POVray³⁰.

four EF hands are involved directly in the formation of the binding pocket. The overall aequorin conformation is most similar to the homologous sarcoplasmic calcium-binding proteins, which form similar four-EF-hand globular structures^{10,11}.

The binding cavity for coelenterazine is situated in the centre of the protein and has a volume of $\sim 600 \text{ \AA}^3$ (Fig. 2). The cavity is closed to a spherical probe of 1.4 Å radius, indicating that there is no solvent access to the site from the surface in this conformation. Lining the cavity are part or all of the side chains of 21 hydrophobic residues, 3 tyrosines, 3 histidines, 2 threonines and 1 lysine. The N ζ group of the sole lysine in the cavity (Lys 39) is involved in protein interactions and only the hydrophobic methylenes of the side chain face towards the cavity. The histidines and tyrosines are organized in three pairs, each closely associated with a tryptophan. These three Tyr-His-Trp triads are notable features that may have functional importance, although their specific roles are unclear at present.

Table 1 Interhelical angles in EF hands of aequorin and some related proteins

	Aequorin	SCBP	Recoverin (1)	Recoverin (2)	Calmodulin (1)	Calmodulin (2)
EF hand	(Ca ²⁺ -free)	(+Ca ²⁺)	(Ca ²⁺ -free)	(+Ca ²⁺)	(Ca ²⁺ -free)	(+Ca ²⁺)
I	70.1	60.1	169.9	113.4*	133.9	89.2
II	108.9	84.2*	135.5	124.5	128.5	87.4
III	111.7	108.0	117.5	92.6	132.7	101.9
IV	105.1	99.9	47.26	46.37*	136.1	94.8

Interhelical angles were calculated using the program interhix (K. Yap, University of Toronto). Values are based on structures in the following Protein Data Bank entries: SCBP (sarcoplasmic calcium-binding protein), 2SCP.pdb; recoverin (1), 1IKU.pdb; recoverin (2), 1JSA.pdb; calmodulin (1), 1CFC.pdb; calmodulin (2), 1CLL.pdb.

* No calcium is bound at this EF hand in this structure.

Table 2 Data collection and refinement statistics for aequorin

Data Collection				
Data set	Native	Se-Met edge	Se-Met peak	Se-Met remote
Wavelength (Å)	1.072	0.97873	0.97861	0.97703
Resolution range (Å)	50–2.3	50–3.5	50–3.5	50–3.5
Measured reflections	241,571	73,921	80,054	74,963
Unique reflections	23,847	7,631	7,652	7,679
Completeness (%)	95.2 (73.6)	99.8 (100)	99.8 (100)	99.8 (100)
overall (final shell)				
$\langle I/\sigma \rangle$ (final shell)	24.7 (6.1)	8.9 (6.4)	7.6 (6.2)	8.0 (5.9)
R_{merge} (%) (final shell)	6.8 (37.0)	10.0 (13.2)	10.1 (13.7)	9.8 (13.5)
Refinement statistics				
Resolution (all data)	50–2.3			
R_{free}	25.3%			
R_{cryst}	21.9%			
Atoms per AU				
Protein (2 mol per AU)	2,992			
Coelenterazine	68			
Water	173			
Average B (Å ²)				
Main chain	33.5			
Side chain	34.9			
Coelenterazine	34.1			
Water	39.0			
R.m.s. deviation				
Bonds (Å)	0.007			
Angles (°)	1.20			

AU, asymmetric unit.

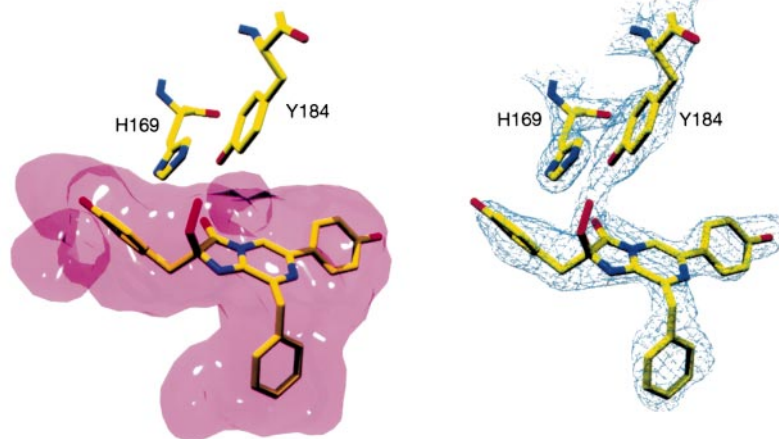


Figure 2 Model of peroxidized coelenterazine with Tyr 184 and His 169. **a**, The peroxidized coelenterazine is shown contained within the binding cavity, which is represented by a pink transparent surface. This surface is the boundary defined by the centre of a 1.4 Å radius spherical probe rolled over the van der Waals surface of the

interior of the cavity in the absence of ligand. **b**, Model with $2F_o - F_c$ electron density map. This map is unbiased by the coelenterazine model, which was omitted from both molecules A and B for low-temperature simulated annealing followed by map calculation. The map is contoured at 1σ .

The luminescence reaction of coelenterazine (a substituted imidazopyrazinone) requires molecular oxygen. In aequorin, both oxygen and coelenterazine are bound³. The addition of calcium results in a conformational transition in the protein that triggers an intramolecular reaction resulting in light emission. The state of the coelenterazine and the oxygen in aequorin has been the subject of some discussion^{3,6,12}. On the basis of several lines of evidence, coelenterazine has been proposed to exist in a peroxidized form, with the peroxide group attached at the C2 position of the ligand^{3,12}. In the present aequorin structure, the coelenterazine is readily recognizable in electron density maps as a continuous multilobed density, even in 3.5 Å maps obtained from the selenomethionyl aequorin crystals. In the native protein at 2.3 Å resolution, the density is well resolved and is consistent with coelenterazine coupled to a hydroperoxide at the C2 position. The resulting chiral C2 is in the *S* configuration (Figs 2 and 3). The hydroperoxide appears to be stabilized by hydrogen bonding to the phenolic oxygen of Tyr 184 which is itself hydrogen bonded to the Ne2 of His 169. The imidazole of His 169, in turn, is situated close to the carbonyl oxygen of C3 on the ligand and to the indole of Trp 173 (Fig. 4). The importance of these interactions is supported by studies that showed that H169F, H169A, H169W and W173F mutants all have little luminescence activity^{13,14}.

In addition to the interactions of the Y184–H169–W173 triad at the C2–C3 reaction centre on the ligand, three other groups of side-chain interactions in the binding cavity are notable. In part, these interactions may position the coelenterazine in the binding site; however, it seems likely that some at least will have additional functional significance to the mechanism of aequorin action, although these functions remain unknown. The three groups of interactions are: (1) the phenolic OH of Tyr 132 is positioned laterally to the plane of the imidazopyrazinone ring of the ligand and hydrogen bonds to N1 (Figs. 3 and 4). It is also linked through hydrogen bonds by a water molecule to His 58. This histidine imidazole is adjacent to the indole ring of Trp 108 which partly overlays the imidazopyrazinone ring system, coming as close as 3.6 Å (Fig. 4). Emphasizing the potential importance of these interactions, H58F and W108F mutants show little calcium-induced luminescence activity^{13–15}. (2) The *p*-OH of the phenol attached at C6 of coelenterazine is positioned at the centre of a triangle made by the third Tyr–His–Trp triad, consisting of the phenolic oxygen of Tyr 82, the Nδ1 of His 16 and the Ne1 of Trp 86 (Figs 3 and 4). An oxygen atom at this position on the ligand, as an OH or as O[−], is not essential to light emission because replacement

of the OH with NH₂ does not abolish luminescence activity¹⁶. (3) The *p*-OH group on the benzyl substituent at C2 of the ligand is hydrogen bonded to a water molecule which itself interacts with the carbonyl oxygen of Ile 105 and the side-chain oxygen of Thr 166 (Figs 3 and 4). This OH group does not appear to be essential to the function of aequorin, as replacing it with hydrogen or halogens does not significantly affect light-emitting capabilities¹⁷. Aequorins containing these modified forms of the ligand, however, show differences in their sensitivity to calcium and rates of luminescence and regeneration.

A number of coelenterazine analogues have been studied for their ability to combine with apoaequorin to produce 'semi-synthetic' forms^{16–18}. The calcium-responsiveness and light-emitting characteristics of many of these forms have been measured. The aromatic rings attached at the C2 and C6 positions seem essential for function, whereas that at C8 is not. Analogues with a wide range of substituents at the C8 position retain activity. Function is also retained with a linker of methylene, dimethylene or ethylene between C5 and the *ortho* position of the phenyl ring at C6, constraining the phenyl ring to be coplanar with the imidazopyrazinone ring. From our structure it is apparent how the capacious hydrophobic binding pocket (Fig. 2a) can accommodate these and other modifications of the ligand, although the subtleties of the effects on light-emission characteristics and calcium-responsiveness remain elusive.

There are general structural similarities between aequorin and calmodulin, but there are also significant differences. The common pairs of helix–loop–helix secondary structure motifs are retained in aequorin, but there are large differences in the interhelical angles of the pairs of helices from those of calmodulin. The interhelical angles of each EF hand of aequorin are shown in Table 1 compared with those in some other proteins in their apo (calcium-free) and calcium-bound forms. On the basis of observations of other EF-hand calcium-binding proteins, it seems likely that the binding of calcium to aequorin will lead to a shift of the interhelical angles of the EF hands. It is difficult to predict the orientations that will result from these shifts given the unique features of aequorin, including the presence of the bound coelenterazine ligand.

Previous studies have suggested that occupancy of two of the three calcium-binding sites in aequorin may be sufficient to trigger activation^{19,20}. Mutation studies to selectively impair calcium-binding at each functional EF-hand loop (I, III and IV) have also been performed¹⁵. The central glycine in each of the active EF-hand loops was substituted in turn by an arginine. Although the struc-

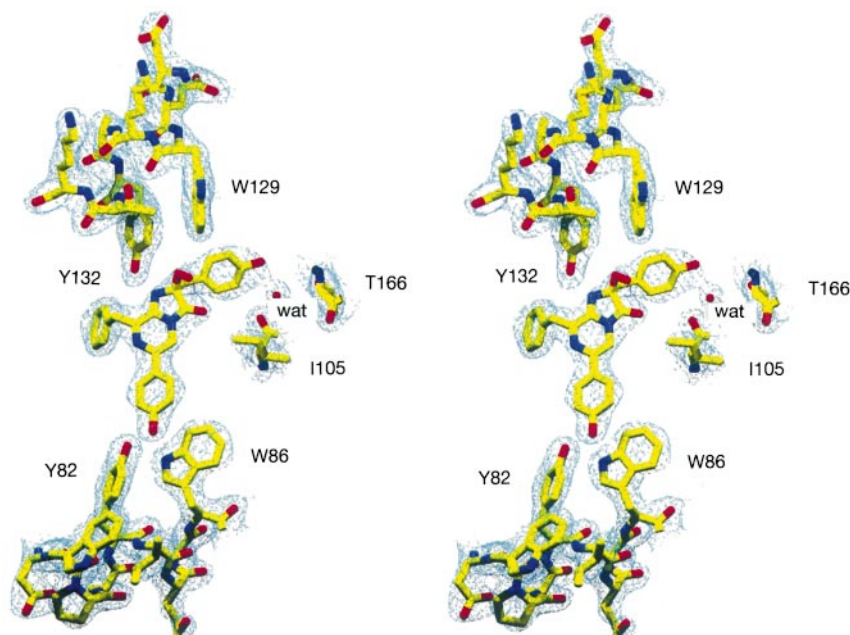


Figure 3 Stereo view of peroxidized coelenterazine (at centre) with some of the surrounding residues. The electron density is a $2F_o - F_c$ map contoured at 1σ .

tural consequences of these mutations are not clear, the mutation in EF hand I led to a complete loss of calcium-induced luminescence, whereas there was little effect with the mutation in domain IV, and an intermediate effect in domain III.

As EF hand II is apparently unable to bind calcium in aequorin, this domain might instead have a role in the enzymic function; however, in our structure there seem to be few special features distinguishing it from the other domains. The two helices form an angle similar to that of the equivalent domain, IV, in the other half of the molecule, and the loop region is remote from the coelenterazine-binding site. The helices form parts of the walls of the ligand-binding cavity and include His 58, Tyr 82 and Trp 86, which, as noted above, at least appear to have roles in positioning the coelenterazine. Despite this, there is no obvious evidence of this region having a preponderance of the residues involved in the coelenterazine oxidation process. As it cannot bind calcium, it is possible that this domain may form a stable scaffold against which the rest of the molecule moves when calcium is bound. Other

interactions may also be possible in this region in the calcium-loaded conformation.

The apparent role of Tyr 184 in stabilizing the coelenterazine peroxide suggests a functional importance to the C-terminal region of the aequorin molecule, which only extends to residue 189. Residues 177–189 form an extended loop structure lying in a space between the first helix of EF hand I and the first helix of EF hand IV (Fig. 1). Numerous hydrogen bonds are formed in this region anchoring the C-terminal chain in its conformation and positioning the side chain of Tyr 184 adjacent to both His 169 and the ligand. Some of these hydrogen bonds also couple the C-terminal chain with both of the flanking helices. This holds the two halves of the molecule together and closes the coelenterazine-binding cavity. The importance of the C terminus has previously been implied by the absolute requirement for a C-terminal proline, with no immediate preceding insertions nor deletions, nor following extensions²¹.

From our present structure, we can consider the consequences of calcium binding by the protein on the conformation of the C-terminal region and how this region may be involved in triggering the light-emission reaction. When calcium binds at either or both EF hands I and IV the helices of these 'hands' would be expected to change their relative orientations, in accordance with observations on other such proteins (Table 1). Displacement of the helices flanking the C-terminal loop would be expected to disrupt the hydrogen-bonding network of the loop, resulting in a relocation of the side chain of Tyr 184. This, in turn, would disrupt the hydrogen bonds to His 169 and the peroxide. No longer stabilized, the peroxide would be free to attack the adjacent carbonylic C3 to initiate the light-emitting reaction. Depending on the extent of the shift in the flanking helices during calcium activation, the C-terminal 'tail' could become partly or completely uncoupled from the helices, opening up the ligand-binding site to permit the egress of one or both reaction products. □

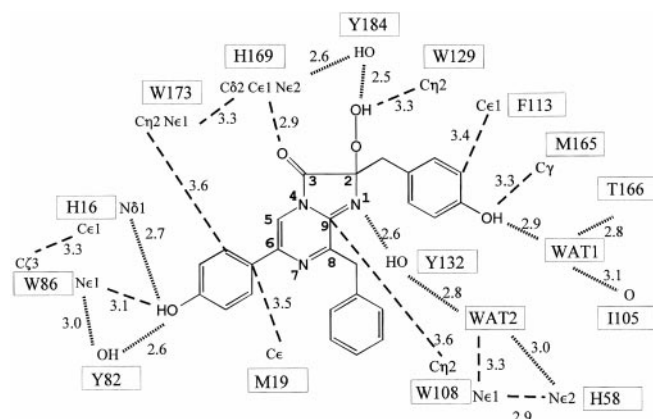


Figure 4 Representation of peroxidized coelenterazine showing all distances to protein atoms within 3.6 Å and some other local interactions. Distances are in ångströms and represent the mean of the distances measured in both A and B molecules. Hydrogen bonds are indicated by dotted lines, other distances are indicated with dashed lines. Numerals around the imidazopyrazinone ring system represent the atom numbering referred to in the text.

Methods

Protein preparation

Recombinant aequorin was prepared as described^{8,22}. Apoaequorin containing selenomethionine was expressed in *Escherichia coli* strain B834(DE3)pLysS (Novagen) carrying plasmid pIP-HE for periplasmic expression. The culture medium was 600 ml of M9 Basic Medium (Sigma) supplemented with 2 mM $MgSO_4$, 0.2% glucose, 0.1 mM $CaCl_2$,

0.01 mM thiamine, 0.004% each of 18 amino acids (the standard ones minus Cys and Met), 0.005% L-(+)-selenomethionine and 30 mg of ampicillin. The seed culture, grown at 30 °C overnight in Luria broth (20 ml), was centrifuged and the pellets added to the culture medium. The mixture was incubated at 36 °C for 16 h with orbital shaking. The bacterial suspension was centrifuged at 12,000g at 0 °C for 8 min. The pellets and the supernatant contained 18 mg and 3 mg of apoaequorin, respectively. The apoaequorin in the pellets was extracted, regenerated into aequorin and purified as described⁸ to yield 14 mg of pure selenomethionine-substituted aequorin. The selenium content of the protein determined by electrospray ionization mass spectrometry indicated that more than 90% of the five methionine residues were substituted.

Crystallization

Crystals were grown at 17 °C by hanging droplet vapour diffusion. Drops consisted of 3 µl of 10 mg ml⁻¹ protein in 2 mM EDTA, 1.2 M ammonium sulphate, 10 mM Bis-Tris pH 7.2 with an equal volume of the well liquor, 63% saturated ammonium sulphate, 10 mM Bis-Tris pH 7.2. Crystals took a week to appear and up to two months to grow to a size suitable for diffraction. The crystals grow as square sectioned rods, the native recombinant protein reaching dimensions of 0.6 × 0.1 × 0.1 mm, the selenomethionyl protein growing no bigger than 0.3 × 0.07 × 0.07 mm.

Data collection

Data were collected at the Brookhaven National Synchrotron Light Source, beamline X8-C. Crystals were cryoprotected by adding glycerol to mother liquor to 15% (v/v). Crystals were flash frozen in the nitrogen stream of an Oxford Cryosystem before data collection. Crystals of native aequorin diffracted to 2.3 Å, whereas smaller selenomethionine-containing crystals diffracted to only 3.5 Å. All data were processed and reduced using DENZO and SCALEPACK²³. The space group of crystals of both protein forms was determined to be P4₃2₁2. Unit-cell dimensions for the native protein were $a = b = 81.27$ Å, $c = 163.66$ Å, for the selenomethionyl protein $a = b = 81.98$ Å, $c = 164.32$ Å. Data were collected on the selenium-containing crystal at three wavelengths to enable MAD phasing (Table 2).

Phasing

Phases were determined from the MAD data using the program SOLVE²⁴. This gave 10 selenium sites, consistent with 2 molecules in the asymmetric unit, 5 sites in each molecule. Initial electron density maps indicated α -helices and molecular boundaries. The phased selenium data were then subjected to density modification using the program DM²⁵, including twofold NCS averaging. Maps obtained from the DM-modified data showed clear electron density for most of the chain of each molecule, including significant side-chain density. On the basis of these maps and the known location of the selenium atom sites, the chain was readily traced from residues 3 to 189 using the program O²⁶.

Refinement

The initial model was refined against the selenium data (peak wavelength) with strict NCS constraints using CNS²⁷ including simulated annealing to 4,000 K. After some rebuilding, this model was subsequently refined using native data, with phases gradually extended using DM to 2.3 Å. To this point, the refined model contained no coelenterazine. Tautomeric models of coelenterazine were built and minimized. An excellent fit to density was obtained using the model based on the tautomer shown in Fig. 3 with a tetragonal, sp^3 , carbon at the C2 position. The exception to the good fit was an unoccupied extension of the density from the position of the C2 atom, strongly emphasized in $F_o - F_c$ maps. This density could be readily fit by a hydroperoxide bonded at C2, as previously suggested^{3,12}. The ligand was therefore rebuilt with the peroxide and used for subsequent modelling. Parameter and topology files for refinement of the peroxidized coelenterazine were obtained using the Hetero Compound Information Centre²⁸. The whole structure was then further refined, including restrained NCS, and water molecules added (Table 2).

As electron density for the peroxide group in the refined model appeared weaker than the rest of the coelenterazine, the peroxide atoms were subjected to occupancy refinement. This showed ~60% occupancy of this part of the ligand in molecule A and ~55% in molecule B. Measurement of light emission from redissolved aequorin crystals shows that when the crystals are obtained from the mother liquor they retain ~90% of the light-emitting capacity of the native protein; however, after X-ray irradiation crystals show less than 50% activity. This is consistent with loss of the peroxide during the data collection as the result of radiation damage.

Received 16 December 1999; accepted 20 March 2000.

- Shimomura, O., Johnson, F. H. & Saiga, Y. Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusa, *Aequorea*. *J. Cell. Comp. Physiol.* **59**, 223–240 (1962).
- Inouye, S. *et al.* Cloning and sequence analysis of cDNA for the luminescent protein aequorin. *Proc. Natl Acad. Sci.* **82**, 3145–3158 (1985).
- Shimomura, O. & Johnson, F. H. Peroxidized coelenterazine, the active group in the photoprotein aequorin. *Proc. Natl Acad. Sci. USA* **75**, 2611–2615 (1978).
- Shimomura, O. & Johnson, F. H. Chemical nature of light emitter in bioluminescence of aequorin. *Tetrahedron Lett.* **31**, 2963–2966 (1973).
- Shimomura, O. & Johnson, F. H. Regeneration of the photoprotein aequorin. *Nature* **256**, 236–238 (1975).
- Prasher, D., McCann, R. O. & Cormier, M. J. Cloning and expression of the cDNA coding for aequorin, a bioluminescent calcium-binding protein. *Biochem. Biophys. Res. Commun.* **126**, 1259–1263 (1985).
- Hannick, L. I., Prasher, D. C., Schultz, L. W., Deschamps, J. R. & Ward, K. B. Preparation and initial characterization of crystals of the photoprotein aequorin from *Aequorea victoria*. *Proteins Struct. Funct. Genet.* **15**, 103–107 (1993).

- Shimomura, O. & Inouye, S. The *in situ* regeneration and extraction of recombinant aequorin from *Escherichia coli* cells and the purification of extracted aequorin. *Protein Expr. Purif.* **16**, 91–95 (1999).
- Tanaka, T., Ames, J. B., Harvey, T. S., Stryer, L. & Ikura, M. Sequestration of the membrane-targeting myristoyl group of recoverin in the calcium-free state. *Nature* **376**, 444–447 (1995).
- Vijay-Kumar, S. & Cook, W. J. Structure of a sarcoplasmic calcium-binding protein from Nereis diversicolor refined at 2.0 Å resolution. *J. Mol. Biol.* **224**, 413–426 (1992).
- Cook, W. J., Jeffrey, L. C., Cox, J. A. & Vijay-Kumar, S. Structure of a sarcoplasmic calcium-binding protein from amphioxus refined at 2.4 Å resolution. *J. Mol. Biol.* **229**, 461–471 (1993).
- Musicki, B., Kishi, Y. & Shimomura, O. Structure of functional part of photoprotein aequorin. *J. Chem. Soc. Chem. Commun.* **126**, 1256–1268 (1986).
- Ohmiya, Y. & Tsuji, F. I. Bioluminescence of the Ca²⁺-binding photoprotein, aequorin, after histidine modification. *FEBS Lett.* **320**, 267–270 (1993).
- Ohmiya, Y., Ohashi, M. & Tsuji, F. I. Two excited states in aequorin bioluminescence induced by tryptophan modification. *FEBS Lett.* **301**, 197–201 (1992).
- Tsuji, F. I., Inouye, S., Goto, T. & Sakaki, Y. Site-specific mutagenesis of the calcium-binding photoprotein aequorin. *Proc. Natl Acad. Sci.* **83**, 8107–8111 (1986).
- Shimomura, O., Musicki, B. & Kishi, Y. Semi-synthetic aequorin. An improved tool for the measurement of calcium ion concentration. *Biochem. J.* **251**, 405–410 (1988).
- Shimomura, O., Musicki, B. & Kishi, Y. Semi-synthetic aequorins with improved sensitivity to Ca²⁺ ions. *Biochem. J.* **261**, 913–920 (1989).
- Shimomura, O., Inouye, S., Musicki, B. & Kishi, Y. Recombinant aequorin and recombinant semi-synthetic aequorins. Cellular Ca²⁺ ion indicators. *Biochem. J.* **270**, 309–312 (1990).
- Shimomura, O. Luminescence of aequorin is triggered by the binding of two calcium ions. *Biochem. Biophys. Res. Commun.* **211**, 359–363 (1995).
- Shimomura, O. & Inouye, S. Titration of recombinant aequorin with calcium chloride. *Biochem. Biophys. Res. Commun.* **221**, 77–81 (1996).
- Nomura, M., Inouye, S., Ohmiya, Y. & Tsuji, F. I. A C-terminal proline is required for bioluminescence of the Ca²⁺-binding photoprotein, aequorin. *FEBS Lett.* **295**, 63–66 (1991).
- Inouye, S., Aoyama, S., Miyata, T., Tsuji, F. I. & Sakaki, Y. Overexpression and purification of the recombinant Ca²⁺-binding protein, apoaequorin. *J. Biochem. (Tokyo)* **105**, 473–477 (1989).
- Otwiński, Z. in *Proceedings of the CCP4 Study Weekend: Data Collection and Processing*. (eds Sawyers, L. Isaacs, N. & Bailey, S.) 56–62 (SERC Daresbury Laboratory, Warrington, UK; 1993).
- Terwilliger, T. C. & Berendzen, J. Automated structure solution for MIR and MAD. *Acta Crystallogr. D* **55**, 849–861 (1999).
- Collaborative Computational Project No. 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Kleywegt, G. J. & Jones, T. A. Databases in protein crystallography. *Acta Crystallogr. D* **54**, 1119–1131 (1998).
- Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723 (1997).
- POV-Ray. Persistence of Vision Raytracer version 3.1. (<http://www.povray.org>)

Acknowledgements

We thank B. Kaminer for initiating this collaborative effort; B. Seaton for involvement in early crystallization studies; H. Nakamura for structural information on imidazopyrazinone; and the Boston University Mass Spectrometry Facility. This work was supported in part by NSF grants to O.S. and to J.F.H.

Correspondence and requests for materials should be addressed to J.F.H. (e-mail: jfh@medxtal.bu.edu). The coordinates of the structure have been deposited in the Protein Data Bank under accession code 1EJ3.

erratum

The ecological cost of sex

C. Patrick Doncaster, Graeme E. Pound & Simon J. Cox

Nature **404**, 281–285 (2000)

In Fig. 2a, the right-hand label on the horizontal axis should have read

$$\frac{k_2}{\alpha_{21}}$$

and the axis label on the horizontal axis should have been 'Abundance of Predator-1 (N_1)'. □

Catching the bug

A selection favouring microbiology — autoclaves, culture flasks and such.

De-ionized water

Purite www.purite.com

Updated pure water range for laboratories

Purite's laboratory de-ionizer cylinder now has colour-coded head fittings allowing identification by type, more efficient transit, faster cylinder change and operation at up to 70 °C. The cylinders are designed for use in laboratories doing general analytical work and glassware washing. The water is free of most mineral impurities as a result of ion exchange. Typical quality is in the region of $10 \text{ M}\Omega \text{ cm}^{-1}$ with output rates of between 10 and 1,000 litres. Purite will handle the entire regeneration process, and can usually reactivate exhausted resin and deliver to site in less than 48 hours.

Reader Service No. 100



Water feature: Purite de-ionized water on tap.

easy-access mouth and clear 25-ml graduations. The polypropylene plug seal caps can be left open to allow gas exchange or closed for a liquid-tight seal. Optional vented caps with a 2- μm hydrophobic membrane are available. The flasks can be used as culture flasks or storage vessels and in shaker and suspension culture applications. They are also recommended for application where biohazardous or radioactive materials make cleaning difficult.

Reader Service No. 102

Easyplate

Bel-Art www.bel-art.com

Stack and carry

The Scienceware Easyplate Petri dish racking system allows one-handed transport of up to 84 100-mm dishes in a single rack. Silicone edges on the vertical channels permit random entry and retrieval with no disturbance to adjacent portions of the stack. Dishes are inserted into the rack media-side up. A dish or group of dishes can then be removed and the area above remains securely in place. The chemically resistant polycarbonate racks are autoclavable and available in a choice of four colours and in both 48-dish and 84-dish sizes.

Reader Service No. 103

ProtoCOL

Synbiosis www.synbiosis.com

Multiple choice

Analysis of bacterial colonies or viral plaques on multi-well plates is no longer limited to manual grading with the ProtoCOL system of automated counting and sizing. This automated colony counter scans each well, counts the number of colonies, measures the area covered by growth/no growth, and calculates the percentage area. Colony counts and area data are delivered within seconds.

Reader Service No. 104

RiboPrinter

Qualicon www.qualicon.com

Molecular typing at your fingertips

The automated RiboPrinter system generates genetic 'fingerprints' of bacteria, allowing molecular typing to be performed in any laboratory, and providing an automated, standardized system for use in existing typing laboratories. Definitive microbial information beyond the species level is available in 8 hours. A new batch of eight samples can be started every two hours. An identification database of RiboPrint patterns based on over 15,000 isolates is included, and Qualicon regularly issues libraries of RiboPrint patterns for additional organisms. Custom libraries can also be created for other genera.

Reader Service No. 105

Alert

Neogen www.neogen.com

Micro-well ELISAs for pathogens

Many micro-well tests for *E. coli* require 90 minutes after enrichment. Alert test kits are now available in a new format with antibodies that reduce this to 35 minutes after enrichment. Combined with Reveal 8-hour media, Alert for *E. coli* 0157 allows testers to go from sample to result in nine hours. Similarly, by eliminating the plating steps involved in conventional laboratory testing for *Campylobacter*, Neogen claim that the Alert kit can detect this pathogen 50 minutes after enrichment.

Reader Service No. 106

Minitron

Infors www.infors.ch

Baby incubator

This 'mini' incubator offers options for cooling, humidification, light manifolds and a gassing system. The magnetic drive and associated electronics are housed in a closed compartment out of the way of spillages of culture medium. The platform pulls out to rest on the

SKR filter cartridges

Meissner www.meissner.com

Small-flow sterilization

Sterilizing grade SKR filter cartridges with Ultradyne PTFE membranes are designed for the sterilization of air and gases in process and product streams, and tank vents. A skirt-flange adapter prevents bypass from upstream to downstream when the cartridge is installed and the housing fully assembled. Meissner claims that these compact cartridges are ideal for new small-flow or retrofit applications to remove fine particulates, bacteria and aerosol viruses from gas lines and sterile tank vents.

Reader Service No. 101

Erlenmeyer flasks

Corning Costar www.corning.com

Open wide

Corning now offer disposable Erlenmeyer flasks in 1- and 2-litre sizes, as well as a 3 litre Fernbach model, in addition to the existing 125, 250 and 500ml models. Made from polycarbonate for maximum optical clarity and mechanical strength, the flasks feature a wide,



Throw-away line: Corning's disposable flasks.



Minitron: small but perfectly formed?

new on the market

door of the chamber for easy access. Cultures shaking in the incubation chamber are visible through an inspection window. A second, independent unit can be stacked above the first to double the shaker capacity without increasing the footprint of 80 × 67 cm.

Reader Service No. 107

Catalogue 2000

New Brunswick Scientific www.nbsc.com

Pick and mix

Hundreds of biological shakers, rockers, rollers, fermenters, bioreactors, continuous flow centrifuges, sterilizers, plate pourers and accessories can be found in New Brunswick's 2000 product catalogue. Ultra-low temperature freezers are also featured. Available from New Brunswick Scientific, Edison House, 163 Dixons Hill Road, North Mymms, Hatfield, Herts AL9 7JE, UK

Reader Service No. 108

Cyclopore

Whatman www.whatman.co.uk

Non-spill membrane filter

The transfer of membranes used for the culture of *E. coli* 0157 and other bacteria from one growth media can be precarious. Hence the new Cyclopore membrane with an outer supporting ring. As well as improving operator safety,

the support ring is designed to prevent curling. Made from track-etched polycarbonate, the uniform 0.4-µm pore size of the new membrane ensures bacteria only grow on its translucent flat surface, to simplify observation under the microscope. The inert nature of polycarbonate means levels of leachables are virtually negligible and will not compromise successful culturing of microorganisms. The filter is available in packs of 25 and has an overall diameter of 82 mm.

Reader Service No. 109

Kestrel

LTE Scientific www.lte-scientific.co.uk

Vacuum packed

A vacuum-assisted option has been added to the Kestrel range of autoclaves. Standard chamber sizes are 200 and 315 litres, and other capacities can be supplied on request. Pre-sterilization vacuum removes air pockets, ensuring effective sterilization of porous loads and rapid attainment of sterilizing temperature. Post-sterilization, the vacuum assists cooling and removes much of the steam and water from the chamber. When programmed for vacuum drying, removal of the remaining residual water from the load is achieved. Steam is kept on constant standby, further accelerating the cycle.

Reader Service No. 110

Status autoclaves

Philip Harris Scientific www.philipharris.co.uk

Dual voltage, greater safety

Status autoclaves have low loading heights and feature assisted lid-lift and safe-to-touch exterior lid covers. They operate from a normal 15-amp supply, a push-button selection enables the faster heating capability of a 30-amp supply to be used. A forced-air cooling module can be fitted which can reduce cooling time by up to as 30 minutes. Variable, timed free-steaming periods can be selected to displace trapped air from the load. A cooling lock ensures that bottled media is at a safe temperature before the lid is released, and can be overridden for immediate venting. An optional printer is available to produce record of sterilisation cycles. The autoclaves are available in 75-litre or 125-litre top-loading versions.

Reader Service No. 111

Yeast deletion mutants

ATCC www.atcc.org

It's a knock-out

The products of the Saccharomyces Genome Deletion Project (SGDP) and Yeast Genetic Stock Center are now available from the ATCC (American Type Culture Collection). The SGDP will enable the analysis of each yeast gene, continuing the study of functional genomics that began after the full sequencing of the yeast genome. There are more than 6,000 open reading frames in the genome. Each deletion mutant will be in a and alpha haploids and in homozygous and heterozygous diploids, resulting in 24,000 knock-out mutants. The on-line search engine can be accessed at www.deletion.stanford.edu or via the ATCC website.

Reader Service No. 112

GMS for fungi assay

BioGenex www.biogenex.com

All that glitters

The Grocott methenamine-silver for fungi (GMS) special stains kit for use on the Opti-Max Pus consolidated staining system contains all reagents necessary for the assay. GMS stains are intended for demonstrating fungi in formalin-fixed, paraffin-embedded surgical pathology, archive or autopsy tissue sections. Only substances that possess large quantities of polysaccharides, such as fungal cell walls, glycogen, and mucins will remain reactive with the silver, reducing it to visible metallic silver. Gold chloride is a toning solution and sodium thiosulphate removes any unreduced silver.

Reader Service No. 113

These notes are compiled in the Nature office from information provided by the manufacturers. For more details, fill in the reader service card.

ADVERTISEMENTS

REAGENT KITS

NEW

FLUOROSPARK PTP KIT
Simple, Rapid and Sensitive
Quantifies PTP activity

ESTROGEN RECEPTOR KIT
High throughput screening of endocrine
disrupting chemicals

SLP REAGENT KIT
Highly sensitive peptidoglycan and
(1-3)-β-glucan Assay

Wako Wako USA Wako GmbH
877-714-1921 (49)02131 3110
www.wako-chem.co.jp



For Immunology Apoptosis,
& Cytokines, order your Y2K
Immuno Kontakt Catalog

UK Tel+44 1235 828200 Fax +44 1235 820482
D Tel+49 69 779099 Fax +49 69 13376880
CH Tel+41 91 6045522 Fax +41 91 6051785
amsbio@diel.pipex.com
www.immunokontakt.com

READER ENQUIRY NO. 41

READER ENQUIRY NO. 42

TREVIGEN

new

PARP

**Poly (ADP-ribose)
Polymerase Reagents**

- PARP inhibition assay
- Biotinylated NAD
- Recombinant human PARP
- Anti PARP and PAR

www.trevigen.com

US and Canada: 1 800 873 8443
TEL: 301 216 2800 • FAX: 301 216 2801

READER ENQUIRY NO. 43

**Custom Peptides &
Polyclonal Antibodies**

Please contact us about complete
services for your custom peptide
and polyclonal antibody needs.

Research Genetics
U.S. or Canada 800-533-4363
FAX 256-536-9016

Check out our homepage at
<http://www.resgen.com>

READER ENQUIRY NO. 39